Информатика и её применения

Том 15 Выпуск 2 Год 2021

СОДЕРЖАНИЕ

Информатика и её применения

Том 15 Выпуск 2 Год 2021

СОДЕРЖАНИЕ

Технический редактор *Л. Кокушкина* Художественный редактор *М. Седакова*
Сдано в набор 04.06.21. Подписано в печать 21.06.21. Формат 60 x 84 / 8
Бумага офсетная. Печать цифровая. Усл.-печ. л. 14,88. Уч.-изд. л. 12,0. Тираж 100 экз.
Заказ № 1135
Издательство «ТОРУС ПРЕСС», Москва 121614, ул. Крылатская, 29-1-43
Отпечатано в НИПКЦ «Восход-А» с готовых файлов
Москва 123103, наб. Новикова-Прибоя, д. 3, корп. 2

УПРАВЛЕНИЕ ЛИНЕЙНЫМ ВЫХОДОМ МАРКОВСКОЙ ЦЕПИ ПО КВАДРАТИЧНОМУ КРИТЕРИЮ*

А. В. Босов[1]

Аннотация: Решена задача оптимального управления выходом стохастической системы наблюдения, в которой состояние определяет ненаблюдаемый марковский скачкообразный процесс, а линейные наблюдения задаются системой дифференциальных уравнений Ито с винеровским процессом. В наблюдения аддитивно входит управление, так что формируется управляемый выход системы. Цель оптимизации задается квадратичным критерием общего вида. Для решения задачи управления сформулирована теорема разделения, использующая решение задачи оптимальной фильтрации, обеспечиваемое фильтром Вонэма. В результате разделения формируется эквивалентная задача управления выходом диффузионного процесса частного вида, а именно: с линейным сносом и нелинейной диффузией. Решение этой задачи обеспечивается непосредственным применением метода динамического программирования.

Ключевые слова: марковский скачкообразный процесс; стохастическая дифференциальная система Ито; оптимальное управление; квадратичный критерий; стохастическая фильтрация; фильтр Вонэма

DOI: 10.14357/19922264210201

1 Введение

В теории управления есть результаты, обладающие особой выразительностью, определяющие не частные решения, а принципы и концепции. К ним, безусловно, относится LQG-задача (LQG — linear-quadratic-Gaussian) — управление линейно-гауссовской стохастической системой по квадратичному критерию [1]. Для целей данной статьи особо важен результат LQG-управления в постановке с неполной информацией, известный как теорема разделения задач управления и фильтрации состояния [2]. Этот результат породил, по-видимому, наиболее действенный подход к синтезу управлений в нелинейных системах наблюдения, который называют принципом разделения.

Считается, что классическими методами, равно как и основанными на них приближенными методами поиска оптимальных управлений [3–5], хорошо решаются задачи с полной информацией о состоянии. Но более востребованными для практики представляются именно задачи с неполной информацией, постановки для систем наблюдения, прежде всего стохастических. Общая теория для таких систем и соответствующих задач основана на уравнении Дункана–Мортенсена–Закаи, описывающего эволюцию апостериорной плотности вероятности и уравнение динамического программирования в вариационных производных [6], развивался этот подход в [7–9]. Но получить оптимальные решения удается крайне редко. Один из примеров дает результат, полученный в [10] для модели управляемой марковской цепи, там же можно оценить сопровождающие такие решения технические трудности.

Ясно, что в задачах с неполной информацией ключевую роль играют методы стохастической фильтрации, поскольку результатами фильтрации заменяют значения фазовых координат в синтезированных по полной информации управлениях. Такая замена и есть принцип разделения, но если в LQG-задаче его реализация обеспечивается формальной теоремой разделения и приводит к оптимальному решению, то в общем случае объединение решений задач управления и фильтрации носит интуитивный характер, а вопрос о потерях качества при постулировании разделения остается открытым. Такие задачи, как в [10], единичны, и это обстоятельство делает ценными любые результаты в задачах оптимизации нелинейных стохастических систем по неполной информации о состоянии. Примеру такой задачи посвящена настоящая работа.

2 Постановка задачи управления выходом цепи

На каноническом вероятностном пространстве $(\Omega, \mathcal{F}, \mathcal{P}, \mathcal{F}_t)$, $t \in [0, T]$, рассмотрим стохастическую

*Работа выполнена при частичной поддержке РФФИ (проект 19-07-00187-А).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, ABosov@frccsc.ru

систему наблюдения с вектором состояния y_t и связанным с ним линейно управляемым выходом $z_t^{(0)}$ (верхний индекс здесь и в других обозначениях введен для удобства дальнейших обозначений):

$$dy_t = \Lambda_t^{\mathrm{T}} y_t\,dt + d\Lambda_t^y\,,\quad y_0 = Y\,; \tag{1}$$

$$dz_t^{(0)} = a_t^{(0)} y_t\,dt + b_t^{(0)} z_t^{(0)} dt + c_t^{(0)} u_t\,dt + \sigma_t^{(0)} dw_t\,,$$

$$z_0^{(0)} = Z^{(0)}\,. \tag{2}$$

Уравнение (1) определяет марковский скачкообразный процесс — цепь с конечным числом состояний и значениями в множестве $\{e_1,\ldots,e_{n_y}\}$, состоящем из единичных координатных векторов в евклидовом пространстве $\mathbb{R}^{n_y}$. Предполагается, что начальное состояние Y имеет известное распределение π, Λ_t — матрица интенсивностей переходов и Λ_t^y — $\mathcal{F}_t$-согласованный мартингал с квадратичной характеристикой [11]

$$\langle \Lambda^y, \Lambda^y\rangle_t = $$

$$= \int_0^t \left(\operatorname{diag}\left(\Lambda_s^{\mathrm{T}} y_s\right) - \Lambda_s^{\mathrm{T}} \operatorname{diag}\left(y_s\right) - \operatorname{diag}\left(y_s\right)\Lambda_s\right) ds.$$

Уравнение (2) представляют косвенные наблюдения $z_t^{(0)} \in \mathbb{R}^{n_z}$ за состоянием цепи y_t и одновременно линейный управляемый выход системы. Порождаемую наблюдениями σ-алгебру будем обозначать $\mathcal{F}_t^{z^{(0)}}$ и предполагать, что

$$\mathcal{F}_t^{z^{(0)}} \subseteq \mathcal{F}_t \subseteq \mathcal{F}\,.$$

Далее, здесь $w_t \in \mathbb{R}^{n_w}$ — не зависящий от Λ_t^y, Y и $Z^{(0)}$ стандартный векторный винеровский процесс; $Z^{(0)} \in \mathbb{R}^{n_z}$ — гауссовская случайная величина с известными моментами, не зависящая от Λ_t^y и Y; $u_t \in \mathbb{R}^{n_u}$, $u_t = U_t(z_t^{(0)})$, $U_t = U_t(z)$, $z \in \mathbb{R}^{n_z}$, — допустимое управление с обратной связью [3]. Для использования выхода $z_t^{(0)}$ в качестве наблюдений будем предполагать невырожденность ошибок наблюдений, т. е. $\sigma_t^{(0)}(\sigma_t^{(0)})^{\mathrm{T}} > 0$.

Качество управления

$$U_0^{\mathrm{T}} = \{U_t(z), 0 \le t \le T\}$$

определяется целевым функционалом следующего вида:

$$J\left(U_0^{\mathrm{T}}\right) = \mathbb{E}\left\{\int_0^{\mathrm{T}} \left\|P_t y_t + Q_t^{(0)} z_t^{(0)} + R_t u_t\right\|_{S_t}^2 dt + \right.$$

$$\left. + \left\|P_T y_T + Q_T^{(0)} z_T^{(0)}\right\|_{S_T}^2\right\}, \tag{3}$$

где $u_t = U_t(z_t^{(0)})$; $P_t \in \mathbb{R}^{n_J\times n_y}$; $Q_t^{(0)} \in \mathbb{R}^{n_J\times n_z}$; $R_t \in \mathbb{R}^{n_J\times n_u}$; $S_t \in \mathbb{R}^{n_J\times n_J}$, $S_t \ge 0$, $S_t = S_t^{\mathrm{T}}$, $0 \le t \le T$, — заданные ограниченные матричные функции. Весовая функция $\|x\|_S^2 = x^{\mathrm{T}} S x$ для симметричной неотрицательно определенной матрицы S, единичной матрице $S = \mathbf{1}$ соответствует евклидова норма $\|x\|_1^2 = |x|^2$ Для исключения возможности отсутствия штрафа для отдельных компонентов вектора управления u_t, которая делает целевой функционал (3) физически некорректным, предполагается выполненным обычное условие невырожденности, в данных обозначениях принимающее вид:

$$R_t^{\mathrm{T}} S_t R_t > 0\,.$$

Функционал (3) отражает одну из традиционных задач автоматического управления — регулирование выхода или управление по выходу. В [12] в качестве практического приложения такой постановки рассматривается задача «доведения выхода до нуля». В других практических целях функционал (3) можно использовать, например, для формализации задачи отслеживания выходом состояния $|y_t - z_t^{(0)}|^2$ или управлением — выхода $|z_t^{(0)} - u_t|^2$, учитывая при этом расходы на управляющее воздействие $|u_t|^2$ и/или значение выходной переменной $|z_t^{(0)}|^2$.

Матричные функции $a_t^{(0)} \in \mathbb{R}^{n_z\times n_y}$, $b_t^{(0)} \in \in \mathbb{R}^{n_z\times n_z}$, $c_t^{(0)} \in \mathbb{R}^{n_z\times n_u}$ и $\sigma_t^{(0)} \in \mathbb{R}^{n_z\times n_w}$ предполагаются ограниченными:

$$\left|a_t^{(0)}\right| + \left|b_t^{(0)}\right| + \left|c_t^{(0)}\right| + \left|\sigma_t^{(0)}\right| \le C$$

для всех $0 \le t \le T$. Таким образом обеспечивается существование решения уравнения (2) для любого допустимого управления u_t. Более того, будем предполагать, что все используемые функции времени Λ_t, $a_t^{(0)}$, $b_t^{(0)}$, $c_t^{(0)}$, $\sigma_t^{(0)}$, P_t, $Q_t^{(0)}$, R_t и S_t кусочно-непрерывны, чтобы обеспечить выполнение типовых условий существования решений обыкновенных дифференциальных уравнений, получаемых далее.

Задачу составляет поиск u_t^* — допустимого управления, доставляющего минимум квадратичному функционалу $J(U_0^{\mathrm{T}})$:

$$(U^*)_0^{\mathrm{T}} = \{U_t^*(z), 0 \le t \le T\} \in \operatorname{argmin} J(U_0^{\mathrm{T}})\,,$$

в предположении существования этого минимума.

3 Разделение задач управления и фильтрации

Для решения поставленной задачи потребуется выполнить некоторые преобразования, которые

позволят показать наличие в рассматриваемой задаче двух важных свойств, необходимых для успешного решения. Заметим, что для системы наблюдения (1), (2) известно решение задачи фильтрации, а именно: уравнение для условного математического ожидания $\mathbb{E}\{y_t|\mathcal{F}_t^{z^{(0)}}\}$, описываемое фильтром Вонэма [11]. Однако записать этот фильтр здесь требуется так, чтобы показать отсутствие в задаче управления дуального эффекта, т. е. влияния выбора оценки состояния на качество управления [13]. Возможно, точнее эту первую промежуточную цель выполняемых преобразований выразит предложение показать независимость качества оптимальной фильтрации от реализуемого закона управления. Вторая цель — преобразовать целевой функционал (3) так, чтобы были выделены независимые слагаемые, определяющие по отдельности качество управления и качество оценивания.

Вначале предлагается выполнить замену переменных в (2), избавляясь от слагаемых $b_t^{(0)} z_t^{(0)} dt$ и $c_t^{(0)} u_t\, dt$. Для такой замены обозначим через $B_t \in \in \mathbb{R}^{n_z \times n_z}$ решение уравнения

$$dB_t = -B_t b_t^{(0)} dt\,,$$

т. е. матричную экспоненту

$$B_t = \exp\left\{ -\int_0^t b_s^{(0)} ds \right\},$$

и через $z_t^{(1)} \in \mathbb{R}^{n_z}$ линейное преобразование выхода

$$z_t^{(1)} = B_t z_t^{(0)} - \int_0^t B_s c_s^{(0)} u_s\, ds\,.$$

Далее, дифференцируя, получаем:

$$\begin{aligned} dz_t^{(1)} &= -B_t b_t^{(0)} z_t^{(0)} dt + \\ &+ B_t \left(a_t^{(0)} y_t\, dt + b_t^{(0)} z_t^{(0)} dt + c_t^{(0)} u_t\, dt + \sigma_t^{(0)} dw_t \right) - \\ &- B_t c_t^{(0)} u_t\, dt = B_t a_t^{(0)} y_t\, dt + B_t \sigma_t^{(0)} dw_t\,, \end{aligned}$$

или

$$dz_t^{(1)} = a_t^{(1)} y_t\, dt + \sigma_t^{(1)} dw_t \qquad (4)$$

с дополнительными обозначениями $a_t^{(1)} = B_t a_t^{(0)}$, $\sigma_t^{(1)} = B_t \sigma_t^{(0)}$. Выполненные преобразования при этом не влияют на решение задачи фильтрации в том смысле, что

$$\mathbb{E}\left\{ y_t|\mathcal{F}_t^{z^{(0)}} \right\} = \mathbb{E}\left\{ y_t|\mathcal{F}_t^{z^{(1)}} \right\},$$

поскольку использованная для получения $z_t^{(1)}$ замена является линейным невырожденным преобразованием $z_t^{(0)}$. Таким образом, оценка оптимального фильтра не зависит от u_t, и в качестве наблюдений можно использовать $z_t^{(1)}$, получая одну и ту же оценку состояния для любого допустимого управления. Соответственно, вместо задачи оценивания состояния по наблюдениям $z_t^{(0)}$, зависящим от реализуемого закона управления, можно рассматривать эквивалентную задачу оценивания y_t по наблюдениям $z_t^{(1)}$, описываемым уравнением (4) и не зависящим от u_t.

Обозначив оптимальную оценку

$$Y_t = \mathbb{E}\left\{ y_t|\mathcal{F}_t^{z^{(1)}} \right\},$$

запишем уравнение для нее, т. е. фильтр Вонэма [11]:

$$\begin{aligned} dY_t = \Lambda_t^{\mathrm{T}} Y_t\, dt + \left(\mathrm{diag}\,(Y_t) - Y_t Y_t^{\mathrm{T}}\right) \left(a_t^{(1)}\right)^{\mathrm{T}} \times \\ \times \left(\sigma_t^{(1)} \left(\sigma_t^{(1)}\right)^{\mathrm{T}}\right)^{-1/2} \times \\ \times \left(\sigma_t^{(1)} \left(\sigma_t^{(1)}\right)^{\mathrm{T}}\right)^{-1/2} \left(dz_t^{(1)} - a_t^{(1)} Y_t\, dt\right), \\ Y_0 = \mathbb{E}\{Y\}\,. \end{aligned} \qquad (5)$$

Сделать это позволяет предположение о невырожденности $\sigma_t^{(1)} (\sigma_t^{(1)})^{\mathrm{T}} > 0$.

Однако использовать уравнение (5) в качестве уравнения состояния пока нельзя, так как требуется вторая замена, нужная для преобразования целевого функционала (3). Эта замена состоит во введении дополнительной переменной в вектор наблюдений, а именно: через $z_t^{(2)} \in \mathbb{R}^{n_z}$ обозначим $\int_0^t B_s c_s u_s\, ds$, т. е. дополним модель наблюдений уравнением

$$dz_t^{(2)} = B_t c_t u_t\, dt\,, \quad z_0^{(2)} = 0\,.$$

Новый вектор наблюдений обозначим

$$Z_t = \begin{pmatrix} z_t^{(1)} \\ z_t^{(2)} \end{pmatrix} \in \mathbb{R}^{nz}, \;\; n_Z = 2n_z\,.$$

Ясно, что такая замена ничего не изменит в отношении оценки оптимального фильтра, т. е.

$$\mathbb{E}\left\{ y_t|\mathcal{F}_t^{z^{(0)}} \right\} = \mathbb{E}\left\{ y_t \,\middle|\, \mathcal{F}_t^{z^{(1)}} \right\} = \mathbb{E}\left\{ y_t|\mathcal{F}_t^{z^{(1)},z^{(2)}} \right\}.$$

Уравнение для Z_t имеет вид:

$$dZ_t = \begin{pmatrix} a_t^{(1)} \\ \mathbf{0} \end{pmatrix} y_t\, dt + \begin{pmatrix} \mathbf{0} \\ B_t c_t \end{pmatrix} u_t\, dt + \begin{pmatrix} \sigma_t^{(1)} \\ \mathbf{0} \end{pmatrix} dw_t\,. \qquad (6)$$

Обозначая блочные матрицы

$$a_t = \begin{pmatrix} a_t^{(1)} \\ \mathbf{0} \end{pmatrix}; \quad c_t = \begin{pmatrix} \mathbf{0} \\ B_t c_t \end{pmatrix}; \quad \sigma_t^{(2)} = \begin{pmatrix} \sigma_t^{(1)} \\ \mathbf{0} \end{pmatrix},$$

где $\mathbf{0}$ — нулевая матрица подходящей размерности, перепишем уравнение (6):

$$dZ_t = a_t y_t\,dt + c_t u_t\,dt + \sigma_t^{(2)} dw_t\,, \tag{7}$$

причем заметим, что

$$z_t^{(0)} = B_t^{-1}\left(z_t^{(1)} + \int\limits_0^t B_s c_s^{(0)} u_s\,ds\right) = \\ = B_t^{-1}\left(z_t^{(1)} + z_t^{(2)}\right) = B_t^{-1}(\mathbf{1}\,\mathbf{1})Z_t\,, \tag{8}$$

где использовано обозначение $(\mathbf{1}\ \mathbf{1})$ для блочной матрицы, составленной из двух единичных $\mathbf{1} \in \in \mathbb{R}^{n_z \times n_z}$. Здесь использован тот факт, что матричная экспонента B_t невырожденная [14], что позволяет применять обратную матрицу B_t^{-1}. Полученное равенство (8) далее используется для замены переменных в целевом функционале (3).

Теперь можно получить уравнение для оптимальной оценки

$$Y_t = \mathbb{E}\left\{y_t | \mathcal{F}_t^Z\right\},$$

используя уравнения (5) и (7). Записать фильтр Вонэма непосредственно для наблюдений Z_t нельзя, поскольку $\sigma_t^{(2)}\left(\sigma_t^{(2)}\right)^{\mathrm{T}}$ вырождена, но вопрос легко решается, поскольку запись (7) носит технический характер, а фактически нужно использовать наблюдения (4), которые легко получить из (7) с помощью блочной матрицы $I = (\mathbf{1}\ \mathbf{0})$, составленной из единичной и нулевой матриц $\mathbf{1}, \mathbf{0} \in \mathbb{R}^{n_z \times n_z}$. С учетом $\sigma_t^{(0)}\left(\sigma_t^{(0)}\right)^{\mathrm{T}} > 0$ и $\sigma_t^{(1)} = I\sigma_t^{(2)} = B_t\sigma_t^{(0)}$ для Y_t имеем:

$$dY_t = \Lambda_t^{\mathrm{T}} Y_t\,dt + \left(\mathrm{diag}\,(Y_t) - Y_t Y_t^{\mathrm{T}}\right)(Ia_t)^{\mathrm{T}} \times \\ \times \left(I\sigma_t^{(2)}\left(I\sigma_t^{(2)}\right)^{\mathrm{T}}\right)^{-1/2} \times \\ \times \left(I\sigma_t^{(2)}\left(I\sigma_t^{(2)}\right)^{\mathrm{T}}\right)^{-1/2} I\left(dZ_t - a_t Y_t\,dt\right), \\ Y_0 = \mathbb{E}\{Y\},$$

где $dW_t = (I\sigma_t^{(2)}(I\sigma_t^{(2)})^{\mathrm{T}})^{-1/2} I(dZ_t - a_t Y_t\,dt)$ определяет $\mathcal{F}_t^Z$-согласованный стандартный векторный винеровский процесс W_t [11]. Это обстоятельство позволяет переписать уравнение наблюдений (7) в виде:

$$IdZ_t = Ia_t Y_t\,dt + \left(I\sigma_t^{(2)}\left(I\sigma_t^{(2)}\right)^{\mathrm{T}}\right)^{1/2} dW_t\,.$$

Здесь было учтено, что

$$I\left(dZ_t - a_t Y_t\,dt - c_t u_t\,dt\right) = I\left(dZ_t - a_t Y_t\,dt\right).$$

Окончательно, обозначив

$$\sigma_t = I^{\mathrm{T}}\left(I\sigma_t^{(2)}\left(I\sigma_t^{(2)}\right)^{\mathrm{T}}\right)^{1/2}; \\ \Sigma_t = \Sigma_t(Y_t) = \\ = \left(\mathrm{diag}\,(Y_t) - Y_t Y_t^{\mathrm{T}}\right)(Ia_t)^{\mathrm{T}}\left(I\sigma_t^{(2)}\left(I\sigma_t^{(2)}\right)^{\mathrm{T}}\right)^{-1/2},$$

получаем систему управления с полной информацией следующего вида:

$$\left.\begin{aligned} dY_t &= \Lambda_t^{\mathrm{T}} Y_t\,dt + \Sigma_t(Y_t)\,dW_t\,, \quad Y_0 = \mathbb{E}\{Y\}; \\ dZ_t &= a_t Y_t\,dt + c_t u_t\,dt + \sigma_t\,dW_t\,, \\ Z_0 &= Z = \begin{pmatrix} Z^{(0)} \\ \mathbf{0} \end{pmatrix}. \end{aligned}\right\} \tag{9}$$

Для окончательного формирования эквивалентной исходной задачи управления остается записать в выбранных переменных имеющийся целевой функционал (3). Обеспечивает это формула полного математического ожидания [15] и замена переменной $z_t^{(0)}$, заданная в (8):

$$J(U_0^{\mathrm{T}}) = \mathbb{E}\left\{\int\limits_0^{\mathrm{T}} \Big\| P_t\,(y_t - Y_t + Y_t) + \right. \\ + Q_t^{(0)} B_t^{-1}(\mathbf{1}\,\mathbf{1})Z_t + R_t u_t\Big\|_{S_t}^2\,dt + \\ \left. + \Big\| P_T\,(y_T - Y_T + Y_T) + Q_T^{(0)} B_T^{-1}(\mathbf{1}\,\mathbf{1})Z_T\Big\|_{S_T}^2 \right\} = \\ = \mathbb{E}\left\{\int\limits_0^{\mathrm{T}} \|P_t Y_t + Q_t Z_t + R_t u_t\|_{S_t}^2 + \right. \\ + \|P_T Y_T + Q_T Z_T\|_{S_T}^2 + \int\limits_0^{\mathrm{T}} \|P_t\,(y_t - Y_t)\|_{S_t}^2 dt + \\ \left. + \|P_T(y_t - Y_t)\|_{S_T}^2 \right\}, \tag{10}$$

где учтено, что $Y_t = \mathbb{E}\{y_t | \mathcal{F}_t^Z\}$ и дополнительно обозначено

$$Q_t = Q_t^{(0)} B_t^{-1}(\mathbf{1}\,\mathbf{1}).$$

Поскольку последнее слагаемое в выражении (10) не зависит от U_0^{T}, а характеризует только точность оптимальной оценки фильтрации Y_t, то его можно из целевого функционала исключить. Таким образом, новый целевой функционал имеет вид:

$$J\left(U_0^{\mathrm{T}}\right) = \mathbb{E}\left\{\int_0^{\mathrm{T}} \|P_t Y_t + Q_t Z_t + R_t u_t\|^2_{S_t} dt + \right.$$
$$\left. + \|P_T Y_T + Q_T Z_T\|^2_{S_T}\right\}. \quad (11)$$

Формулировка окончательного утверждения такова.

Утверждение 1 (*теорема разделения*). *Решение задачи оптимального управления с полной информацией ($\mathcal{F}_t^{Y,Z}$-измеримого) системой* (9) *с целевым функционалом* (11) *является оптимальным решением задачи управления системой* (1), (2) *с косвенными наблюдениями ($\mathcal{F}_t^{z^{(0)}}$-измеримого) с целевым функционалом* (3).

Заметим, что управление $(U^*)_0^{\mathrm{T}}$, минимизирующее функционал (11), минимизирует и (3), поэтому для целевой функции не имеет смысла вводить новое обозначение.

4 Решение задачи управления выходом

Полученное далее решение основано на подходе, представленном в [16] в задаче управления выходом системой с состоянием, описываемым уравнением Ито, и работе [17], где проанализирован частный случай этой же задачи в случае коррелированных возмущений. Найти решение, как и в [16, 17], сформулированной выше задачи управления выходом Z_t формируемым состоянием Y_t, с квадратичной функцией цены $J(U_0^{\mathrm{T}})$ удается классическим методом динамического программирования [3, 5]. Обозначив функцию Беллмана через

$$V_t = V_t(y,z),\ \ y \in \mathbb{R}^{n_y},\ z \in \mathbb{R}^{n_z},\ \Sigma_t = \Sigma_t(y),$$

получим уравнение динамического программирования

$$\frac{\partial V_t}{\partial t} +$$
$$+\frac{1}{2}\operatorname{tr}\left\{\Sigma_t^{\mathrm{T}}\frac{\partial^2 V_t}{\partial y^2}\Sigma_t + \sigma_t^{\mathrm{T}}\frac{\partial^2 V_t}{\partial z^2}\sigma_t + 2\Sigma_t^{\mathrm{T}}\frac{\partial^2 V_t}{\partial y \partial x}\sigma_t\right\} +$$
$$+\min_u\left\{y^{\mathrm{T}}\Lambda_t\frac{\partial V_t}{\partial y} + (a_t y + c_t u)^{\mathrm{T}}\frac{\partial V_t}{\partial z} + \right.$$
$$\left. + \|P_t y + Q_t z + R_t u\|^2_{S_t}\right\} = 0,$$
$$V_T = \|P_T y + Q_T z\|^2_{S_T}. \quad (12)$$

Существование решения уравнения (12) является достаточным условием оптимальности, оптимальное управление при этом — точка минимума соответствующего слагаемого. Нетрудно видеть, что при сделанном выше предположении $R_t^{\mathrm{T}} S_t R_t > > 0$ этот минимум доставляет оптимальное управление (в предположении существования решения (12)):

$$u_t^* = u_t^*(y,z) = -\frac{1}{2}\left(R_t^{\mathrm{T}} S_t R_t\right)^{-1}\left(c_t^{\mathrm{T}}\frac{\partial V_t}{\partial z} + \right.$$
$$\left. + 2R_t^{\mathrm{T}} S_t\left(P_t y + Q_t z\right)\right). \quad (13)$$

Подставляя u_t^* в (12) и перегруппировывая слагаемые, получаем:

$$\frac{\partial V_t}{\partial t} + \frac{1}{2}\operatorname{tr}\left\{\Sigma_t^{\mathrm{T}}\frac{\partial^2 V_t}{\partial y^2}\Sigma_t + \sigma_t^{\mathrm{T}}\frac{\partial^2 V_t}{\partial z^2}\sigma_t + \right.$$
$$\left. + 2\Sigma_t^{\mathrm{T}}\frac{\partial^2 V_t}{\partial y \partial z}\sigma_t\right\} + y^{\mathrm{T}}\Lambda_t\frac{\partial V_t}{\partial y} + \Big(y^{\mathrm{T}} a_t^{\mathrm{T}} -$$
$$- \left(R_t^{\mathrm{T}} S_t\left(P_t y + Q_t z\right)\right)^{\mathrm{T}}\left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} c_t^{\mathrm{T}}\Big)\frac{\partial V_t}{\partial z} + (P_t y +$$
$$+ Q_t z)^{\mathrm{T}}\left(S_t - S_t R_t\left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} R_t^{\mathrm{T}} S_t\right)(P_t y +$$
$$+ Q_t z) - \frac{1}{4}\left(\frac{\partial V_t}{\partial z}\right)^{\mathrm{T}} c_t\left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} c_t^{\mathrm{T}}\frac{\partial V_t}{\partial z} = 0. \quad (14)$$

Рассматривая полученное уравнение, заметим, что линейное вхождение Z_t в уравнение наблюдений (9) и квадратичное в целевой функционал (11) позволяет предположить, что решение уравнения (14) может быть представлено в виде (векторная форма представления, предложенного в [16]):

$$V_t = V_t(y,z) = z^{\mathrm{T}}\alpha_t z + z^{\mathrm{T}}\beta_t(y) + \gamma_t(y), \quad (15)$$

что сводит поиск оптимального решения к поиску уравнений относительно симметричной матричной функции α_t, векторной функции $\beta_t(y)$ и скалярной функции $\gamma_t(y)$, причем явный вид функции $\gamma_t(y)$ для реализации оптимального управления не требуется, нужна только производная

$$\frac{\partial V_t}{\partial z} = 2\alpha_t z + \beta_t(y).$$

Представление функции Беллмана (15) можно упростить дальше, используя то, что слагаемое с производной $\partial V_t/\partial y$ в (14) содержит только множитель $y^{\mathrm{T}}\Lambda_t$. Это позволяет предположить, что $\beta_t(y)$ является аффинным преобразованием y (адаптированная к рассматриваемой задаче форма, предложенная в [17]):

$$\beta_t(y) = \beta_t^{MN} y\,, \tag{16}$$

где матрица $\beta_t^{MN} \in \mathbb{R}^{n_z \times n_y}$. Выбранное здесь обозначение β_t^{MN} будет ясно из дальнейшего (см. далее (21), где уравнение для β_t^{MN} определяется матрицами M_t^{β} и N_t^{β}). Граничное условие при этом принимает вид:

$$V_T = \|P_T y + Q_T z\|^2_{S_T} =$$
$$= z^{\mathrm{T}} Q_T^{\mathrm{T}} S_T Q_T z + 2 z^{\mathrm{T}} Q_T^{\mathrm{T}} S_T P_T y + y^{\mathrm{T}} P_T^{\mathrm{T}} S_T P_T y\,,$$

т. е.

$$\left.\begin{aligned} \alpha_T &= Q_T^{\mathrm{T}} S_T Q_T\,; \\ \beta_T^{MN} &= 2 Q_T^{\mathrm{T}} S_T P_T\,; \\ \gamma_T(y) &= y^{\mathrm{T}} P_T^{\mathrm{T}} S_T P_T y\,, \end{aligned}\right\} \tag{17}$$

а оптимальное управление (13) окончательно:

$$u_t^* = u_t^*(y,z) =$$
$$= -\frac{1}{2}\left(R_t^{\mathrm{T}} S_t R_t\right)^{-1}\left(c_t^{\mathrm{T}}\left(2\alpha_t z + \beta_t^M y\right) + \right.$$
$$\left. + 2 R_t^{\mathrm{T}} S_t \left(P_t y + Q_t z\right)\right)\,.$$

Таким образом, u_t^* содержит два слагаемых: первый терм, линейный по z, т. е. по наблюдениям, точнее по переменной выхода; второй терм, линейный по y, т. е. по состоянию, точнее по оценке состояния, заданной фильтром Вонэма. Чтобы сказанное было верно, нужно получить уравнения для α_t, β_t^{MN} и $\gamma_t(y)$, показав, что сделанное предположение относительно функции V_t с учетом аффинности $\beta_t(y)$ позволяет решить уравнение (14). Для этого подставляем (15) в уравнение (14) и учитываем, что производные $\beta_t(y)$ согласно (16) равны

$$\frac{\partial^2 \beta_t(y)}{\partial y^2} = 0\,; \quad \frac{\partial \beta_t(y)}{\partial y} = \left(\beta_t^{MN}\right)^{\mathrm{T}}\,,$$

а также что

$$y^{\mathrm{T}}\Lambda_t \frac{\partial V_t}{\partial y} = \left(\frac{\partial V_t}{\partial y}\right)^{\mathrm{T}} \Lambda_t^{\mathrm{T}} y = z^{\mathrm{T}} \beta_t^{MN} \Lambda_t^{\mathrm{T}} y + y^{\mathrm{T}} \Lambda_t \frac{\partial \gamma_t}{\partial y};$$
$$\frac{\partial^2 V_t}{\partial y \partial z} = \left(\beta_t^{MN}\right)^{\mathrm{T}}\,.$$

После этой подстановки и незначительных преобразований получаются уравнения для α_t — коэффициента при z^{T} и z, $\beta_t(y) = \beta_t^{MN} y$ — коэффициента при z^{T} и y, а также для $\gamma_t = \gamma_t(y)$ — оставшихся слагаемых — функций y:

$$\frac{d\alpha_t}{dt} - \left(M_t^{\alpha}\alpha_t + \alpha_t^{\mathrm{T}}(M_t^{\alpha})^{\mathrm{T}}\right) + N_t^{\alpha} -$$
$$- \alpha_t^{\mathrm{T}} c_t \left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} c_t^{\mathrm{T}} \alpha_t = 0\,; \tag{18}$$

$$\frac{d\beta_t^{MN}}{dt}\, y + \beta_t^{MN}\Lambda_t^{\mathrm{T}} y + M_t^{\beta} y - N_t^{\beta}\beta_t^{MN} y = 0\,; \tag{19}$$

$$\frac{\partial \gamma_t}{\partial t} + \frac{1}{2}\mathrm{tr}\left\{\Sigma_t^{\mathrm{T}} \frac{\partial^2 \gamma_t}{\partial y^2}\,\Sigma_t\right\} +$$
$$+ \mathrm{tr}\left\{\sigma_t^{\mathrm{T}}\alpha_t\sigma_t + \Sigma_t^{\mathrm{T}}\left(\beta_t^{MN}\right)^{\mathrm{T}}\sigma_t\right\} +$$
$$+ y^{\mathrm{T}}\Lambda_t \frac{\partial \gamma_t}{\partial y} + M_t^{\gamma} = 0\,, \tag{20}$$

где обозначено

$$M_t^{\alpha} = Q_t^{\mathrm{T}} S_t R_t \left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} c_t^{\mathrm{T}}\,;$$
$$N_t^{\alpha} = Q_t^{\mathrm{T}}\left(S_t - S_t R_t \left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} R_t^{\mathrm{T}} S_t\right) Q_t\,;$$
$$M_t^{\beta} = 2\left(\left(a_t^{\mathrm{T}} - P_t^{\mathrm{T}} S_t R_t \left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} c_t^{\mathrm{T}}\right)\alpha_t + \right.$$
$$\left. + P_t^{\mathrm{T}}\left(S_t - S_t R_t \left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} R_t^{\mathrm{T}} S_t\right) Q_t\right)\,;$$
$$N_t^{\beta} = Q_t^{\mathrm{T}} S_t R_t \left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} c_t^{\mathrm{T}} +$$
$$+ \alpha_t c_t \left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} c_t^{\mathrm{T}}\,;$$
$$M_t^{\gamma} = M_t^{\gamma}(y) = \beta_t^{\mathrm{T}}\left(a_t - c_t\left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} R_t^{\mathrm{T}} S_t P_t\right) y +$$
$$+ y^{\mathrm{T}} P_t^{\mathrm{T}}\left(S_t - S_t R_t \left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} R_t^{\mathrm{T}} S_t\right) P_t y -$$
$$- \frac{1}{4}\beta_t^{\mathrm{T}} c_t \left(R_t^{\mathrm{T}} S_t R_t\right)^{-1} c_t^{\mathrm{T}}\beta_t\,.$$

Наконец, из (19) получаются уравнения для β_t^{MN}:

$$\frac{d\beta_t^{MN}}{dt} + \beta_t^{MN}\Lambda_t^{\mathrm{T}} + M_t^{\beta} - N_t^{\beta}\beta_t^{MN} = 0\,. \tag{21}$$

Решаются полученные уравнения с граничными условиями (17). При этом уравнение (18) — это матричное уравнением Риккати для квадратной симметричной матрицы α_t. Сделанных выше предположений в отношении кусочной непрерывности коэффициентов этого уравнения и условия $R_t^{\mathrm{T}} S_t R_t > 0$ достаточно для существования единственного неотрицательного решения для всех $0 \le t \le T$. Действительно, такое уравнение имеется в классической линейно-квадратичной задаче и, как известно [18], существует единственное оптимальное управление — линейное с обратной связью

по выходу Z_t, с коэффициентом усиления, описываемым этим уравнением Риккати. Соотношения (21) представляют собой систему обыкновенных линейных дифференциальных уравнений относительно элементов матрицы β_t^{MN} с кусочно-непрерывными коэффициентами, так что существование и единственность решения обеспечиваются обычной для линейных дифференциальных уравнений теоремой. Уравнение (20) для $\gamma_t(y)$ является линейным дифференциальным уравнением в частных производных параболического типа. Предполагается существование решения у этого уравнения, и это предположение интерпретируется как достаточное условие, такое же как использованное уравнение Беллмана (12). Подводит итог рассуждениям раздела следующее утверждение.

Утверждение 2 (*управление выходом для случая полной информации*). *Если существует решение уравнения динамического программирования* (12), *то это решение может быть представлено в виде* (15), (16) *так, что коэффициенты* α_t, $\beta_t(y) = \beta_t^M y$ *и* $\gamma_t(y)$ *определяются уравнениями* (18), (21) и (20) *соответственно, а оптимальное управление*

$$u_t^* = u_t^*(Y_t, Z_t) = - \\ = \frac{1}{2}\left(R_t^{\mathrm{T}} S_t R_t\right)^{-1}\left(c_t^{\mathrm{T}}\left(2\alpha_t Z_t + \beta_t^{MN} Y_t\right) + \right. \\ \left. + 2R_t^{\mathrm{T}} S_t\left(P_t Y_t + Q_t Z_t\right)\right). \quad (22)$$

5 Заключение

Принципиальный результат статьи представляет, во-первых, найденная явная форма (22) оптимального управления в рассматриваемой задаче, во-вторых, тот факт, что это оптимальное решение оформлено в виде управления с обратной связью, в котором ключевую роль играет решение вспомогательной задачи оптимальной фильтрации — условное математическое ожидание $Y_t = \mathbb{E}\{y_t|\mathcal{F}_t^Z\}$. Это и дало основание сформулировать основной результат, назвав его теоремой разделения. Принципиальное отличие полученного решения от классической LQG-задачи состоит в том, что измененная для разделения задача не похожа на исходную, а именно: мартингальное представление марковской цепи (1) заменяется совсем иным объектом — стохастическим уравнением Ито с винеровским процессом (9).

Таким образом, ключом к решению оказалось уникальное свойство фильтра Вонэма, представившего оценку марковской цепи по косвенным, зашумленным гауссовским шумом наблюдениям в виде классического уравнения Ито с винеровским процессом.

Литература

1. *Athans M.* The role and use of the stochastic linear-quadratic-Gaussian problem in control system design // IEEE T. Automat. Contr., 1971. Vol. 16. No. 6. P. 529–552.
2. *Wonham W. M.* On the separation theorem of stochastic control // SIAM J. Control, 1968. Vol. 6. No. 2. P. 312–326.
3. *Флеминг У., Ришел Р.* Оптимальное управление детерминированными и стохастическими системами / Пер. с англ. — М.: Мир, 1978. 316 с. (*Fleming W. H., Rishel R. W.* Deterministic and stochastic optimal control. — New York, NY, USA: Springer-Verlag, 1975. 222 p.).
4. *Kushner H. J., Dupuis P. G.* Numerical methods for stochastic control problems in continuous time. — New York, NY, USA: Springer-Verlag, 2001. 476 p.
5. *Bertsekas D. P.* Dynamic programming and optimal control. — Cambridge: Athena Scientific, 2017. 576 p.
6. *Mortensen R. E.* Stochastic optimal control with noisy observations // Int. J. Control, 1966. Vol. 4. No. 5. P. 455–464.
7. *Davis M. H. A., Varaiya P. P.* Dynamic programming conditions for partially observable stochastic systems // SIAM J. Control, 1973. Vol. 11. No. 2. P. 226–262.
8. *Benes V. E., Karatzas I.* On the relation of Zakai's and Mortensen's equations // SIAM J. Control Optim., 1983. Vol. 21. No. 3. P. 472–489.
9. *Bensoussan A.* Stochastic control of partially observable systems. — Cambridge: Cambridge University Press, 1992. 364 p.
10. *Miller B. M., K. E. Avrachenkov, K. V. Stepanyan, G. B. Miller.* The problem of optimal stochastic data flow control based upon incomplete information // Probl. Inf. Transm., 2005. Vol. 41. No. 2. P. 150–170.
11. *Elliott R. J., Aggoun L., Moore J. B.* Hidden Markov models: Estimation and control. — New York, NY, USA: Springer-Verlag, 1995. 382 p.
12. *Athans M., Falb P. L.* Optimal control: An introduction to the theory and its applications. — New York, NY, USA: Dover Publications, 2007. 879 p.
13. *Фельдбаум А. А.* Основы теории оптимальных автоматических систем. — 2-е изд. — М.: Наука, 1966. 624 с.
14. *Bhatia R.* Matrix analysis. — Graduate texts in mathematics ser. — New York, NY, USA: Springer-Verlag, 1997. Vol. 169. 349 p.
15. *Ширяев А. Н.* Вероятность. — 2-е изд. — М.: Наука, 1989. 640 с.
16. *Босов А. В., Стефанович А. И.* Управление выходом стохастической дифференциальной системы по квадратичному критерию. I. Оптимальное решение

методом динамического программирования // Информатика и её применения, 2018. Т. 12. Вып. 3. С. 99–106.

17. *Босов А. В.* О некоторых частных случаях в задаче управления выходом стохастической дифференциальной системы по квадратичному критерию // Информатика и её применения, 2021. Т. 15. Вып. 1. С. 11–17.
18. *Девис М. Х. А.* Линейное оценивание и стохастическое управление / Пер. с англ. — М.: Наука, 1984. 206 с. (*Davis M. H. A.* Linear estimation and stochastic control. — London: Chapman and Hall, 1977. 224 p.)

Поступила в редакцию 24.12.2020

LINEAR OUTPUT CONTROL OF MARKOV CHAINS BY THE QUADRATIC CRITERION

A. V. Bosov

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The problem of optimal output control of a stochastic observation system, in which the state determines an unobservable Markov jump process and linear observations are given by a system of Ito differential equations with a Wiener process, is solved. Observations additively include control vector, so that a controlled output of the system is formed. The optimization goal is set by a general quadratic criterion. To solve the control problem, a separation theorem is formulated that uses the solution to the optimal filtering problem provided by the Wonham filter. As a result of the separation, an equivalent problem of output control of a diffusion process of a particular type, namely, with linear drift and nonlinear diffusion, is formed. The solution of this problem is provided by direct application of the dynamic programming method.

Keywords: Markov jump process; Ito stochastic differential system; optimal control; quadratic criterion; stochastic filtering; Wonham filter

DOI: 10.14357/19922264210201

Acknowledgments

This work was partially supported by the Russian Foundation for Basic Research (grant 19-07-00187-A).

References

1. Athans, M. 1971. The role and use of the stochastic linear-quadratic-gaussian problem in control system design. *IEEE T. Automat. Contr.* 16(6):529–552.
2. Wonham, W. M. 1968. On the separation theorem of stochastic control. *SIAM J. Control* 6(2):312–326.
3. Fleming, W. H., and R. W. Rishel. 1975. *Deterministic and stochastic optimal control*. New York, NY: Springer-Verlag. 222 p.
4. Kushner, H. J., and P. G. Dupuis. 2001. *Numerical methods for stochastic control problems in continuous time*. New York, NY: Springer-Verlag. 476 p.
5. Bertsekas, D. P. 2017. *Dynamic programming and optimal control*. Cambridge: Athena Scientific. 576 p.
6. Mortensen, R. E. 1966. Stochastic optimal control with noisy observations. *Int. J. Control* 4(5):455–464.
7. Davis, M. H. A., and P. P. Varaiya. 1973. Dynamic programming conditions for partially observable stochastic systems. *SIAM J. Control* 11(2):226–262.
8. Benes, V. E., and I. Karatzas. 1983. On the relation of Zakai's and Mortensen's equations. *SIAM J. Control Optim.* 21(3):472–489.
9. Bensoussan, A. 1992. *Stochastic control of partially observable systems*. Cambridge: Cambridge University Press. 364 p.
10. Miller, B. M., K. E. Avrachenkov, K. V. Stepanyan, and G. B. Miller. 2005. The problem of optimal stochastic data flow control based upon incomplete information. *Probl. Inf. Transm.* 41(2):150–170.
11. Elliott, R. J., L. Aggoun, and J. B. Moore. 1995. *Hidden Markov models: Estimation and control*. New York, NY: Springer-Verlag. 382 p.
12. Athans, M., and P. L. Falb. 2007. *Optimal control: An introduction to the theory and its applications*. New York, NY: Dover Publications. 879 p.
13. Feldbaum, A. A. 1966. *Osnovy teorii optimal'nykh avtomaticheskikh system* [Foundations of theory of optimal automatic systems]. Moscow: Nauka. 624 p.

14. Bhatia, R. 1997. *Matrix analysis*. Graduate texts in mathematics ser. New York, NY: Springer-Verlag. Vol. 169. 349 p.
15. Shiryaev, A. N. 1996. *Probability*. New York, NY: Springer Verlag. 624 p.
16. Bosov, A. V., and A. I. Stefanovich. 2018. Upravlenie vykhodom stokhasticheskoy differentsial'noy sistemy po kvadratichnomu kriteriyu. I. Optimal'noe reshenie metodom dinamicheskogo programmirovaniya [Stochastic differential system output control by the quadratic criterion. I. Dynamic programming optimal solution]. *Informatika i ee Primeneniya — Inform. Appl.* 12(3):99–106.
17. Bosov, A. V. 2021. O nekotorykh chastnykh sluchayakh v zadache upravleniya vykhodom stokhasticheskoy differentsial'noy sistemy po kvadratichnomu kriteriyu [On some special cases in the problem of stochastic differential system output control by the quadratic criterion]. *Informatika i ee Primeneniya — Inform. Appl.* 15(1):11–17.
18. Davis, M. H. A. 1977. *Linear estimation and stochastic control*. London: Chapman and Hall. 224 p.

Received December 24, 2020

Contributor

Bosov Alexey V. (b. 1969) — Doctor of Science in technology, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; AVBosov@ipiran.ru

ФИЛЬТРАЦИЯ СОСТОЯНИЙ МАРКОВСКИХ СКАЧКООБРАЗНЫХ ПРОЦЕССОВ ПО КОМПЛЕКСНЫМ НАБЛЮДЕНИЯМ I: ТОЧНОЕ РЕШЕНИЕ ЗАДАЧИ*

А. В. Борисов[1], Д. Х. Казанчян[2]

Аннотация: Первая часть цикла посвящена решению задачи оптимальной фильтрации состояний марковских скачкообразных процессов (МСП) по совокупности наблюдаемых диффузионных и считающих процессов. Интенсивности шумов в наблюдаемой диффузии и скачков в разрывных наблюдениях зависят от оцениваемого состояния. Предложено специальное эквивалентное преобразование наблюдений, приводящее их к совокупности диффузионного процесса с единичной диффузией, множеству считающих процессов и совокупности косвенных наблюдений, выполненных в неслучайные дискретные моменты времени. Искомая оптимальная оценка представима в форме решения дискретно-непрерывной стохастической дифференциальной системы с преобразованными наблюдениями в правой части. Приведено условие идентифицируемости, при выполнении которого состояние МСП может быть восстановлено по косвенным зашумленным наблюдениям точно.

Ключевые слова: марковский скачкообразный процесс; оптимальная фильтрация; мультипликативные шумы в наблюдениях; непрерывные и считающие наблюдения; условия идентифицируемости

DOI: 10.14357/19922264210202

1 Введение

Данный цикл может рассматриваться как продолжение работ [1–3], посвященных теоретическому и практическому аспектам решения задачи оптимальной фильтрации состояний *марковских скачкообразных процессов* по косвенным зашумленным наблюдениям. На этот раз наблюдения пополнены процессами Кокса, интенсивность которых зависит от оцениваемого состояния МСП. Данная работа посвящена теоретическому решению задачи фильтрации.

Статья организована следующим образом.

Раздел 2 содержит формальную постановку задачи оптимальной фильтрации.

В разд. 3 предложено специальное преобразование наблюдений, позволяющее решить поставленную задачу, а также утверждение, представляющее дискретно-непрерывную стохастическую систему, описывающую искомую оценку.

Заключительные замечания приведены в разд. 4.

2 Постановка задачи фильтрации

На полном вероятностном пространстве с фильтрацией $(\Omega, \mathcal{F}, \mathsf{P}, \{\mathcal{F}_t\}_{t\geqslant 0})$ рассматривается стохастическая динамическая система

$$X_t = X_0 + \int\limits_0^t \Lambda^{\top}(s) X_s \, ds + \mu_t^X; \tag{1}$$

$$Y_t = \int\limits_0^t f(s) X_s \, ds + \int\limits_0^t \sum_{n=1}^N X_s^n g_n^{1/2}(s) \, dW_s; \tag{2}$$

$$Z_t = \int\limits_0^t h(s) X_s \, ds + \mu_t^Z, \tag{3}$$

где

– $X_t \triangleq \mathrm{col}(X_t^1, \ldots, X_t^N) \in \mathbb{S}^N$ — ненаблюдаемое состояние системы — МСП с множеством состояний $\mathbb{S}^N \triangleq \{e_1, \ldots, e_N\}$ ($\mathbb{S}^N$ — множество единичных векторов пространства $\mathbb{R}^N$), матричнозначной функцией интенсивностей пе-

*Работа выполнена при частичной поддержке РФФИ (проект 19-07-00187 А) и в соответствии с программой Московского центра фундаментальной и прикладной математики.

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; Московской авиационный институт; Факультет вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; Центр фундаментальной и прикладной математики Московского государственного университета имени М. В. Ломоносова, ABorisov@frccsc.ru

[2]Факультет вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова, Drastamat94@gmail.com

реходов $\Lambda(t)$ и начальным распределением π; $\mu_t^X \triangleq \mathrm{col}(\mu_t^{X_1}, \ldots, \mu_t^{X_N}) \in \mathbb{R}^N$ — $\mathcal{F}_t$-согласованный мартингал;

– $Y_t \triangleq \mathrm{col}(Y_t^1, \ldots, Y_t^M) \in \mathbb{R}^M$ — косвенные наблюдения, зашумленные $\mathcal{F}_t$-согласованным стандартным винеровским процессом $W_t \triangleq \triangleq \mathrm{col}(W_t^1, \ldots, W_t^M) \in \mathbb{R}^M$; $f(t)$ — $(M \times N)$-мерная матричнозначная функция плана наблюдений, а набор $(M \times M)$-мерных симметричных матричнозначных функций $\{g_n(t)\}_{n=\overline{1,N}}$ характеризует интенсивности шумов в зависимости от текущего состояния X_t;

– $Z_t \triangleq \mathrm{col}(Z_t^1, \ldots, Z_t^K) \in \mathbb{R}^K$ — косвенные наблюдения, компоненты которого являются считающими процессами: элементы $(K \times N)$-мерной матричнозначной функции $h(t)$ определяют интенсивность скачков отдельных компонент в зависимости от текущего состояния X_t; $\mu_t^Z \triangleq \mathrm{col}(\mu_t^{Z_1}, \ldots, \mu_t^{Z_K}) \in \mathbb{R}^N$ — $\mathcal{F}_t$-согласованный мартингал.

Обозначим через $\{\mathcal{O}_t\}$ поток σ-алгебр всех наблюдений, полученных на отрезке времени $[0, t]$, $\mathcal{O}_0 \triangleq \{\varnothing, \Omega\}$, а $\{\overline{\mathcal{O}}_t\}$ ($\mathcal{O}_{t+} \triangleq \cap_{s:s \geqslant t} \mathcal{O}_s$) — вариант данного потока, замкнутый справа.

Задача оптимальной фильтрации состояния МСП X_t заключается в построении *условного математического ожидания* (УМО) $\widehat{X}_t \triangleq \mathsf{E}\left\{X_t | \overline{\mathcal{O}}_t\right\}$.

Ниже представлены ограничения на исследуемую систему наблюдения (1)–(3), необходимые для корректного решения поставленной задачи фильтрации.

А. Поток σ-подалгебр $\mathcal{F}_t$ непрерывен справа. Все траектории МСП $\{X_t\}_{t \geqslant 0}$ непрерывны справа и имеют конечные пределы слева, т. е. являются cádlág-процессами.

Б. Все компоненты $\Lambda(t)$, $f(t)$, $\{g_n(t)\}_{n=\overline{1,N}}$ и $h(t)$ — неслучайные cádlág-процессы.

В. Шумы в Y равномерно невырожденны [4], т. е. для любых $t \geqslant 0$, $1 \leqslant n \leqslant N$ верно неравенство $g_n(t) \geqslant \alpha I > 0$ (здесь и ниже I обозначает единичную матрицу подходящей размерности).

Г. Процессы

$$K_{ij}(t) \triangleq \mathbf{I}_{\{\mathbf{0}\}}\left(g_i(t) - g_j(t)\right), \quad i, j = \overline{1, N}, \quad (4)$$

имеют локально ограниченную вариацию; здесь и ниже $\mathbf{I}_{\mathcal{A}}(x)$ обозначает индикаторную функцию множества $\mathcal{A}$, а $\mathbf{0}$ — нулевую матрицу подходящей размерности.

Д. Компоненты мартингала μ^Z процесса Z (2) ортогональны друг другу, т. е.

$$\langle Z \rangle_t = \int_0^t \mathrm{diag}\left(h(s) X_s\right) ds\,.$$

Мартингалы в процессе состояния μ^X и в считающих наблюдениях μ^Z также ортогональны $\langle X, Z \rangle_t \equiv \mathbf{0}$.

Условия А–В являются стандартными в задачах фильтрации [4–6]. Условие Г обеспечивает регулярность разбиения временно́й оси на множества, где какие-либо функции $\{g_n(\cdot)\}$ совпадают или различны: на любом конечном промежутке времени $[0, t]$ для каждой пары (i, j), $i, j = \overline{1, N}$, множество, на котором интенсивности $g_i(\cdot)$ и $g_j(\cdot)$ совпадают, представимо в виде объединения конечного набора интервалов. В частности, это условие выполнено в случае, когда все компоненты $\{g_n(\cdot)\}$ кусочно-гладкие с локально ограниченными производными [7].

Ограничение Д также не представляется обременительным: ортогональность мартингалов считающих процессов означает, что скачки этих процессов *почти наверное* (п. н.) не совпадают. В случае если мартингалы в наблюдаемых процессах не ортогональны, исходные наблюдения всегда можно трансформировать так, чтобы обеспечить требуемую ортогональность преобразованных наблюдений. Для этого достаточно в считающих наблюдениях синхронные скачки различных компонент выделить в отдельные процессы. Условие ортогональности мартингалов в состоянии и наблюдениях позволяет избежать излишней громоздкости формул.

Рассмотрение задач фильтрации относительно замыкания справа $\{\mathcal{O}_t\}$ имеет техническую подоплеку. Дело в том, что математический аппарат стохастического анализа развит для случая, когда поток σ-алгебр, порожденных наблюдениями, непрерывен справа. В общем случае п. н. непрерывность траекторий процессов не гарантирует непрерывности справа потока σ-алгебр, порожденных этими процессами [8]. В [9] показано, что для наблюдений (2), (3) эта непрерывность нарушается. Переход к $\{\overline{\mathcal{O}}_t\}$ позволяет успешно решить теоретическую задачу оптимальной фильтрации. Данная техническая особенность не коснется практической реализации решения задачи. Во второй части работы будут предложены аппроксимации теоретических оценок, вычисляемые по исходному незамкнутому потоку σ-алгебр наблюдений $\{\mathcal{O}_t\}$.

3 Преобразование непрерывных наблюдений и решение задачи фильтрации

В [10] представлен фундаментальный результат — формула УМО специального семимартингала

по наблюдениям также в форме специального семимартингала. В общем случае эта формула не позволяет задать искомое УМО в виде решения конечной замкнутой системы стохастических уравнений. Ситуация упрощается, если наблюдения с помощью гирсановской замены вероятностной меры сводятся к совокупности винеровских и пуассоновских процессов [11, 12]. Наблюдения (2) и (3) не удовлетворяют этим условиям, однако существует преобразование, приводящее их к эквивалентной совокупности диффузионного процесса с единичной диффузией, набора считающих процессов и наблюдений, выполняемых в неслучайные дискретные моменты времени.

Рассмотрим квадратичную характеристику

$$\langle Y\rangle_t = Y_tY_t^\top - \int\limits_0^t Y_s\,dY_s^\top - \int\limits_0^t dY_sY_s^\top = \int\limits_0^t \sum_{n=1}^N X_s^n g_n(s)\,ds\,.$$

Введем в рассмотрение $\overline{\mathcal{O}}_t$-согласованный процесс

$$q_t \triangleq \frac{d\langle Y\rangle_s}{ds}|_{s=t+} = \sum_{n=1}^N X_t^n g_n(t) = q_t\,(X_t)\,. \qquad (5)$$

Так как на любом конечном отрезке времени процессы X и $\{g_n\}$ п. н. терпят лишь конечное число скачков, то $\overline{\mathcal{O}}_t$-согласованный процесс $U_t \triangleq \triangleq \int_0^t q_s^{-1/2}\,dY_s$ допускает мартингальное разложение

$$U_t = \int\limits_0^t \overline{f}(s)X_s\,ds + \overline{W}_t,$$

где

$$\overline{f}(t) \triangleq \sum_{n=1}^N g_n^{-1/2}(t)f(t)\mathrm{diag}e_n;$$

$\overline{W}_t$ — некоторый $\mathcal{F}_t$-согласованный стандартный винеровский процесс [4]. Преобразуем наблюдения q_t так, чтобы построить полный прообраз преобразования (5) в форме N-мерного случайного процесса H_t:

$$H_t \triangleq \sum_{n=1}^N \mathbf{I}_{\{\mathbf{0}\}}\,(q_t - g_n(t))\,e_n\,.$$

Из А и Б следует, что H_t — cádlág-процесс, представимый в виде:

$$H_t = \sum_{n,k=1}^N \mathbf{I}_{\{\mathbf{0}\}}\,(g_k(t+) - g_n(t+))\,X_t^k e_n = K(t+)X_t\,,$$

где элементы матричнозначной функции $K(\cdot)$ определяются формулой (4). Из определения H_t и $K(t)$ следует, что существует такая матричнозначная функция $T(t) : \mathbb{R}_+ \to \mathbb{R}^{N\times N}$, что процесс $V_t \triangleq T(t)H_t$ обладает следующими свойствами:

(i) все компоненты V_t являются cádlág-процессами, причем $\mathsf{P}\left\{V_t \in \mathbb{S}^N\right\} \equiv 1$;

(ii) матрица $J(t) \triangleq T(t)K(t+)$ состоит из cádlág-элементов, является трапециевидной, а ее строки — ортогональные векторы, сформированные из 0 и 1.

В силу своей структуры процесс V_t допускает разложение

$$V_t = D_t + R_t\,.$$

Здесь

$$D_t = J(0)X_0 + \sum_{\substack{s_j\leqslant t:\\ \Delta J(s_j)\neq\mathbf{0}}} \Delta J(s_j)X_{s_j};\;\; R_t = \int\limits_0^t J(s)\,dX_s\,,$$

где $\Delta J(t) \triangleq J(t) - \Delta J(t-)$ — функция скачков $J(t)$. По сути, процесс D_t аккумулирует скачки V_t, порожденные изменениями матрицы $J(t)$ в неслучайные моменты времени s_i, а R_t — скачки V_t, порожденные переходами состояния X_t в случайные моменты времени τ_j. Заметим, что с вероятностью 1 множества моментов $\{s_i\}$ и $\{\tau_j\}$ не пересекаются. Процесс R_t взаимно однозначно определяется процессом $C_t \triangleq \mathrm{col}(C_t^1, \ldots, C_t^N)$, компоненты которого считают число скачков R_t в состояние e_n, $n = \overline{1,N}$, произошедших за время $[0,t]$:

$$C_t^n = \int\limits_0^t (1 - e_n^\top V_{s-})e_n^\top\,dR_s.$$

Лемма 1. *Для любого $t \geqslant 0$ верно тождество $\sigma\{U_s, Z_s, C_s, D_s :\; 0 \leqslant s \leqslant t\} \equiv \overline{\mathcal{O}}_t$.*

Истинность леммы 1 следует из взаимно однозначного соответствия траекторий Y и (U, C, D):

$$U_t \triangleq \int\limits_0^t q_s^{-1/2}dY_s,\;\; q_t = \frac{d\langle Y\rangle_s}{ds}\Big|_{s=t+};$$

$$C_t = \int\limits_0^t (I - \mathrm{diag}\,V_{s-})\,dV_s - \sum_{\substack{s_j\leqslant t:\\ \Delta J(s_j)\neq\mathbf{0}}} (I - \mathrm{diag}\,V_{s_j-})\Delta V_{s_j};$$

$$D_t = V_0 + \sum_{\substack{s_j\leqslant t:\\ \Delta J(s_j)\neq\mathbf{0}}} \Delta V_{s_j},\;\; V_t = T(t)H_t,$$

где

$$H_t \triangleq \sum_{n=1}^N \mathbf{I}_{\{\mathbf{0}\}}(q_t - g_n(t))e_n;$$

$$V_t = D_t + \int\limits_0^t \sum_{i,j=1}^N V^i_{s-}\,(e_j - e_i)\,dC^j_s;$$

$$Y_t = \int\limits_0^t \sum_{n=1}^N V^n_s g^{1/2}_n(s)\,dU_s.$$

Таким образом исходные наблюдения были преобразованы в совокупность:

- непрерывного диффузионного процесса U_t с единичной диффузией;
- совокупности считающих процессов Z_t и C_t;
- наблюдений D_t, полученных в дискретные неслучайные моменты времени.

Ниже используем обозначения $J_n(t) \triangleq e_n^\top J(t)$ для n-й строки $J(t)$, $h_k(t) \triangleq e_k^\top h(t)$ для k-й строки $h(t)$, а также

$$\Gamma_n(t) \triangleq \operatorname{diag}\left(J_n(t)\right)\Lambda^\top(t)\left(I - \operatorname{diag} J_n(t)\right),\ n = \overline{1,N}.$$

Лемма 2. *Верны следующие утверждения.*

1. *Процессы C^n_t, $n = \overline{1,N}$, допускают следующее мартингальное разложение*:

$$C^n_t = \int\limits_0^t \mathbf{1}\Gamma_n(s)X_s\,ds + \int\limits_0^t \left(1 - J_n(s)X_{s-}\right)J_n(s)\,d\mu^X_s\,.$$

2. $\langle C^n, C^m\rangle_t \equiv 0$ *для любых* $n \neq m$;

$$\langle C^n, C^n\rangle_t = \int\limits_0^t \mathbf{1}\Gamma_n(s)X_s\,ds\,.$$

3. $\langle C^n, Z^m\rangle_t \equiv 0$ *для любых* $n = \overline{1,N}$, $m = \overline{1,M}$.
4. *Обновляющие процессы*

$$\nu^n_t \triangleq \int\limits_0^t \left(dC^n_s - \mathbf{1}\Gamma_n(s)\widehat{X}_s\,ds\right),\ \ n = \overline{1,N};$$

$$\varkappa^k_t \triangleq \int\limits_0^t \left(dZ^k_s - h_k(s)\widehat{X}_s\,ds\right),\ \ k = \overline{1,K},$$

являются $\overline{\mathcal{O}}_t$-согласованными мартингалами с квадратичными характеристиками

$$\left.\begin{aligned}\langle\nu^n\rangle_t &= \int\limits_0^t \mathbf{1}\Gamma_n(s)\widehat{X}_s\,ds\,;\\ \langle\varkappa^k\rangle_t &= \int\limits_0^t h_k(s)\widehat{X}_s\,ds\,,\end{aligned}\right\} \qquad (6)$$

а обновляющий процесс

$$\omega_t \triangleq U_t - \int\limits_0^t \overline{f}(s)\widehat{X}_s\,ds \qquad (7)$$

является $\overline{\mathcal{O}}_t$-согласованным M-мерным стандартным винеровским процессом.

Доказательство леммы 2 подобно доказательству соответствующего утверждения в [9].

Теорема 1. *Оптимальная оценка фильтрации $\widehat{X}_t$ является сильным решением стохастической дифференциальной системы*

$$\begin{aligned}\widehat{X}_t = {}& \left((D_0)^\top J(0)\pi_0\right)^+ \operatorname{diag}\left(D_0\right)J(0)\pi_0 + \\ &+ \int\limits_0^t \Lambda^\top(s)\widehat{X}_s\,ds + \int\limits_0^t \mathbf{k}_s\overline{f}^\top(s)\,d\omega_s + \sum_{n=1}^N \int\limits_0^t \left(\Gamma_n(s) - \right.\\ &\left. - \mathbf{1}\Gamma_n(s)\widehat{X}_{s-}I\right)\widehat{X}_{s-}\left(\mathbf{1}\Gamma_n(s)\widehat{X}_{s-}\right)^+ d\nu^n_s + \\ &+ \sum_{k=1}^K \int\limits_0^t \mathbf{k}_{s-}h^\top_k(s)\left(h_k(s)\widehat{X}_{t-}\right)^+ d\varkappa^k_s + \\ &+ \sum_{\substack{s_j\leqslant t:\\ \Delta J(s_j)\neq\mathbf{0}}} \left(\left(\Delta D^\top_{s_j}\Delta J(s_j)\widehat{X}_{s_j-}\right)^+ \operatorname{diag}\left(\Delta D_{s_j}\right)\times\right.\\ &\left.\times\,\Delta J(s_j) - I\right)\widehat{X}_{s_j-}, \qquad (8)\end{aligned}$$

где

$$\mathbf{k}_t = \operatorname{diag}\left(\widehat{X}_t\right) - \widehat{X}_t\widehat{X}_t^\top, \qquad (9)$$

а A^+ обозначает операцию псевдообращения. Решение системы единственно в классе неотрицательных кусочно-непрерывных $\overline{\mathcal{O}}_t$-согласованных процессов с точками разрыва, принадлежащими множеству точек разрыва процесса V.

Доказательство теоремы 1 приведено в приложении.

Теорема 1 позволяет точно восстанавливать состояние МСП X_t по косвенным зашумленным наблюдениям (2), (3) в случае выполнения условий идентифицируемости, представленных в следующей лемме.

Лемма 3. *Если для любых $n \neq m$ ($n, m = \overline{1,N}$) неравенства $G_n(s) \neq G_m(s)$ верны почти всюду по мере Лебега на $[0,t]$, то $\widehat{X}_t = X_t$ P-п. н., и X_t является решением* (8).

Доказательство леммы 3 аналогично доказательству соответствующего утверждения в [9].

4 Заключение

Первая часть цикла содержит теоретическое решение задачи оптимальной фильтрации по комплексным (непрерывным и считающим) наблюдениям при наличии в непрерывных наблюдениях *мультипликативных шумов*, т. е. шумов, интенсивность которых зависит от оцениваемого состояния. Оценка фильтрации определяется решением некоторой замкнутой дискретно-непрерывной стохастической системы, в правую часть которой входят наблюдения, преобразованные специальным образом. Вид оценки имеет важное теоретическое значение, так как позволяет описать важные свойства оценки, в частности условия полной идентифицируемости состояния, когда его можно восстановить точно, основываясь на косвенных зашумленных наблюдениях. Тем не менее этот вид малопригоден для конструирования численных алгоритмов фильтрации из-за того, что необходимое преобразование наблюдений включает в себя операцию, подразумевающую двойной предельный переход: во-первых, вычисление квадратичной характеристики $\langle Y\rangle$ диффузионных наблюдений и, во-вторых, вычисление ее правой производной. Последующие исследования посвящены эффективным численным методам решения данной задачи фильтрации.

Приложение

Доказательство теоремы 1. Используем подход из [12, Part III, Sect. 8.7]. Из правила Байеса следует, что

$$\widehat{X}_0 = \mathsf{E}\left\{X_0|\mathcal{O}_{0+}\right\} = \\ = \mathsf{E}\left\{X_0|D_0\right\} = \left(D_0^\top J(0)\pi\right)^+ \operatorname{diag}\left(D_0\right) J(0)\pi\,.$$

Пусть s_{j-1} — неслучайный момент $(j-1)$-го дискретного наблюдения $\Delta D_{s_{j-1}}$. Исследуем поведение X_t на интервале $[s_{j-1}, s_j)$:

$$X_t = X_{s_{j-1}} + \int\limits_{s_{j-1}}^{t} \Lambda^\top(s) X_s\, ds + \mu_t^X - \mu_{s_{j-1}}^X.$$

Вычисляя УМО обеих частей равенства относительно $\overline{\mathcal{O}}_t$, можно показать, что

$$\widehat{X}_t = \widehat{X}_{s_{j-1}} + \int\limits_{s_{j-1}}^{t} \Lambda^\top(s)\widehat{X}_s\, ds + \mu_t^1,$$

где $\{\mu_t^1\}_{t\in[s_{j-1},s_j)}$ — $\overline{\mathcal{O}}_t$-согласованный мартингал. Для любого $t\in[s_{j-1},s_j)$ верно равенство

$$\overline{\mathcal{O}}_t = \overline{\mathcal{O}}_{s_{j-1}} \vee \sigma\{U_s, Z_s\ s\in(s_{j-1},t]\} \vee \\ \vee\, \sigma\{C_s^n, s\in(s_{j-1},t],\ \ n=\overline{1,N}\}\,.$$

Процесс $\{\omega_t\}$ (7) является $\overline{\mathcal{O}}_t$-согласованным стандартным винеровским; U_t является $\overline{\mathcal{O}}_t$-согласованным семимартингалом с условно независимыми относительно $\mathcal{F}^X$ приращениями, в то время как $\{C_t^n\}_{n=\overline{1,N}}$ и $\{Z_t^k\}_{k=\overline{1,K}}$ — $\overline{\mathcal{O}}_t$-согласованные точечные процессы. Поэтому мартингал μ_t^1 *допускает интегральное представление* [10, Chap. 4, § 8, Problem 1], т. е.

$$\widehat{X}_t = \widehat{X}_{s_{j-1}} + \int\limits_{s_{j-1}}^{t} \Lambda^\top(s)\widehat{X}_s\, ds + \int\limits_{s_{j-1}}^{t} \alpha_s\, d\omega_s + \\ + \int\limits_{s_{j-1}}^{t} \sum_{n=1}^{N} \beta_s^n\, d\nu_s^n + \int\limits_{s_{j-1}}^{t} \sum_{k=1}^{K} \gamma_s^k\, d\varkappa_s^k, \quad (10)$$

где α_t, $\{\beta_t^n\}_{n=\overline{1,N}}$ и $\{\gamma_t^k\}_{k=\overline{1,K}}$ представляют собой $\overline{\mathcal{O}}_t$-предсказуемые процессы подходящей размерности, подлежащие определению.

По обобщенному правилу Ито

$$X_t U_t^\top = X_{s_{j-1}} U_{s_{j-1}}^\top + \\ + \int\limits_{s_{j-1}}^{t} \left(\Lambda^\top(s) X_s U_s^\top + \operatorname{diag}\left(X_s\right)\overline{f}^\top(s)\right) ds + \mu_t^2,$$

где μ_t^2 — некоторый $\mathcal{F}_t$-согласованный мартингал. Беря УМО от обеих частей равенства относительно $\overline{\mathcal{O}}_t$, можно показать, что

$$\widehat{X}_t U_t^\top = \widehat{X}_{s_{j-1}} U_{s_{j-1}}^\top + \\ + \int\limits_{s_{j-1}}^{t} \left(\Lambda^\top(s)\widehat{X}_s U_s^\top + \operatorname{diag}\left(\widehat{X}_s\right)\overline{f}^\top(s)\right) ds + \mu_t^3, \quad (11)$$

где μ_t^3 — некоторый $\overline{\mathcal{O}}_t$-согласованный мартингал. В то же время из правила Ито, представления (10) и факта, что ω_t — винеровский процесс следует, что

$$\widehat{X}_t U_t^\top = \widehat{X}_{s_{j-1}} U_{s_{j-1}}^\top + \\ + \int\limits_{s_{j-1}}^{t} \left(\Lambda^\top(s)\widehat{X}_s U_s^\top + \widehat{X}_s\widehat{X}_s^\top\overline{f}^\top(s) + \alpha_s\right) ds + \mu_t^4, \quad (12)$$

где μ_t^4 — некоторый $\overline{\mathcal{O}}_t$-согласованный мартингал. Так как (11) и (12) являются двумя разложениями одного и того же специального семимартингала $\widehat{X}_t U_t^\top$, то из его единственности следует, что $\overline{\mathcal{O}}_t$-предсказуемый процесс α_t удовлетворяет равенству

$$\int\limits_{s_{j-1}}^{t} \operatorname{diag}\left(\widehat{X}_s\right)\overline{f}^\top(s) ds = \int\limits_{s_{j-1}}^{t} \left(\widehat{X}_s\widehat{X}_s^\top\overline{f}^\top(s) + \alpha_s\right) ds$$

и с учетом (9) и свойств ω_t может быть выбран в виде $\alpha_t = \mathbf{k}_t\overline{f}^\top(t)$.

Вновь из правила Ито, свойств МСП X и C^n получаем:

$$X_t C_t^n = X_{s_{j-1}} C_{s_{j-1}}^n + \\ + \int_{s_{j-1}}^t \left(\Lambda^\top(s) X_s C_s^n + \Gamma_n(s) X_s\right) ds + \mu_t^5,$$

где μ_t^5 — $\mathcal{F}_t$-согласованный мартингал. Вычисляя УМО относительно $\overline{\mathcal{O}}_t$, получаем:

$$\widehat{X}_t C_t^n = \widehat{X}_{s_{j-1}} C_{s_{j-1}}^n + \\ + \int_{s_{j-1}}^t \left(\Lambda^\top(s) \widehat{X}_s C_s^n + \Gamma_n(s) \widehat{X}_s\right) ds + \mu_t^6, \quad (13)$$

где μ_t^6 — $\overline{\mathcal{O}}_t$-согласованный мартингал.

С другой стороны, по правилу Ито и с учетом (10) и (6) можно показать, что

$$\widehat{X}_t C_t^n = \widehat{X}_{s_{j-1}} C_{s_{j-1}}^n + \\ + \int_{s_{j-1}}^t \left(\Lambda^\top(s) \widehat{X}_s C_s^n + \widehat{X}_s \mathbf{1}\Gamma_n(s) \widehat{X}_s + \beta_s^n \mathbf{1}\Gamma_n(s) \widehat{X}_s\right) ds + \\ + \mu_t^7, \quad (14)$$

где μ_t^7 — $\overline{\mathcal{O}}_t$-согласованный мартингал. Из совпадения разложений (13) и (14) можно заключить, что процесс β_s^n должен удовлетворять равенству

$$\int_{s_{j-1}}^t \Gamma_n(s) \widehat{X}_s \, ds = \int_{s_{j-1}}^t \left[\widehat{X}_s \mathbf{1}\Gamma_n(s) \widehat{X}_s + \beta_s^n \mathbf{1}\Gamma_n(s) \widehat{X}_s\right] ds$$

и может быть выбран в форме

$$\beta_t^n = \left(\Gamma_n(t) - \mathbf{1}\Gamma_n(t) \widehat{X}_{t-} I\right) \widehat{X}_{t-} \left(\mathbf{1}\Gamma_n(t) \widehat{X}_{t-}\right)^+, \\ n = \overline{1, N}.$$

Вновь из правила Ито получаем

$$X_t Z_t^k = X_{s_{j-1}} Z_{s_{j-1}}^k + \\ + \int_{s_{j-1}}^t \left(\Lambda^\top(s) X_s Z_s^k + \operatorname{diag}(X_s) h_k^\top(s)\right) ds + \mu_t^8,$$

где μ_t^8 — $\mathcal{F}_t$-согласованный мартингал. Вычисляя УМО относительно $\overline{\mathcal{O}}_t$, получаем

$$\widehat{X}_t Z_t^k = \widehat{X}_{s_{j-1}} Z_{s_{j-1}}^k + \\ + \int_{s_{j-1}}^t \left(\Lambda^\top(s) \widehat{X}_s Z_s^k + \operatorname{diag}\left(\widehat{X}_s\right) h_k^\top(s)\right) ds + \mu_t^9, \quad (15)$$

где μ_t^9 — $\overline{\mathcal{O}}_t$-согласованный мартингал. С другой стороны, по правилу Ито и результатам леммы 2 можно получить равенство

$$\widehat{X}_t Z_t^k = \widehat{X}_{s_{j-1}} Z_{s_{j-1}}^k + \\ + \int_{s_{j-1}}^t \left(\Lambda^\top(s) \widehat{X}_s Z_s^k + \widehat{X}_s \widehat{X}_s^\top h_k^\top(s) + \gamma_s^k h_k(s) \widehat{X}_s\right) ds + \\ + \mu_t^{10}, \quad (16)$$

где μ_t^{10} — $\overline{\mathcal{O}}_t$-согласованный мартингал. Из совпадения разложений (15) и (16) следует, что процесс γ_s^k должен удовлетворять равенству

$$\int_{s_{j-1}}^t \operatorname{diag}\left(\widehat{X}_s\right) h_k^\top(s) \, ds = \int_{s_{j-1}}^t \left(\widehat{X}_s \widehat{X}_s^\top h_k^\top(s) + \gamma_s^k h_k(s) \widehat{X}_s\right) ds$$

и может быть выбран в форме

$$\gamma_t^k = \mathbf{k}_{t-} h_k^\top(t) \left(h_k(t) \widehat{X}_{t-}\right)^+, \ \ k = \overline{1, K}.$$

Таким образом, на интервале $[s_{j-1}, s_j)$ оптимальная оценка $\widehat{X}_t$ описывается стохастической дифференциальной системой

$$\widehat{X}_t = \widehat{X}_{s_{j-1}} + \int_{s_{j-1}}^t \Lambda^\top(s) \widehat{X}_{s-} \, ds + \int_{s_{j-1}}^t \mathbf{k}_s \overline{f}^\top(s) \, d\omega_s + \\ + \sum_{n=1}^N \int_{s_{j-1}}^t \left(\Gamma_n(s) - \mathbf{1}\Gamma_n(s) \widehat{X}_{s-} I\right) \widehat{X}_{s-} \left(\mathbf{1}\Gamma_n(s) \widehat{X}_{s-}\right)^+ d\nu_s^n + \\ + \sum_{k=1}^K \int_{s_{j-1}}^t \mathbf{k}_{s-} h_k^\top(s) \left(h_k(s) \widehat{X}_{s-}\right)^+ d\varkappa_s^k. \quad (17)$$

Так как $\mathsf{P}\left\{\Delta X_{s_j} = 0\right\} = 1$, то с вероятностью 1

$$\mathsf{E}\left\{X_{s_j} | \overline{\mathcal{O}}_{s_{j-1}} \vee \sigma\left\{U_s, Z_s \ s \in (s_{j-1}, s_j]\right\} \vee \right. \\ \left. \vee \sigma\left\{C_s^n, \ s \in (s_{j-1}, s_j], \ \ n = \overline{1, N}\right\}\right\} = \\ = \widehat{X}_{s_{j-1}} + \int_{s_{j-1}}^{s_j} \Lambda^\top(s) \widehat{X}_{s-} \, ds + \int_{s_{j-1}}^{s_j} \mathbf{k}_s \overline{f}^\top(s) \, d\omega_s + \\ + \sum_{n=1}^N \int_{s_{j-1}}^{s_j} \left(\Gamma_n(s) - \mathbf{1}\Gamma_n(s) \widehat{X}_{s-} I\right) \widehat{X}_{s-} \left(\mathbf{1}\Gamma_n(s) \widehat{X}_{s-}\right)^+ d\nu_s^n + \\ + \sum_{k=1}^K \int_{s_{j-1}}^{s_j} \mathbf{k}_{s-} h_k^\top(s) \left(h_k(s) \widehat{X}_{s-}\right)^+ d\varkappa_s^k = \widehat{X}_{s_j-}.$$

В силу того что

$$\overline{\mathcal{O}}_{s_j} = \overline{\mathcal{O}}_{s_{j-1}} \vee \sigma\left\{U_s, Z_s, \ s \in (s_{j-1}, s_j]\right\} \vee \\ \vee \sigma\{C_s^n, \ s \in (s_{j-1}, s_j], \ \ n = \overline{1, N}\} \vee \sigma\left\{\Delta D_{s_j}\right\},$$

по правилу Байеса получаем

$$\widehat{X}_{s_j} = \\ = \left(\Delta D_{s_j}^\top \Delta J(s_j) \widehat{X}_{s_j-}\right)^+ \operatorname{diag}\left(\Delta D_{s_j}\right) \Delta J(s_j) \widehat{X}_{s_j-}. \quad (18)$$

Уравнение (8) получается путем «склейки» локальных решений (17), описывающих изменение $\widehat{X}_t$ на интервалах $[s_{j-1}, s_j)$, и формулы (18), описывающей пересчет оценок в моменты s_j поступления дискретных наблюдений ΔD.

Единственность сильного решения в классе неотрицательных кусочно-непрерывных $\overline{\mathcal{O}}_t$-согласованных процессов с моментами скачков, принадлежащими множеству скачков V_t, доказывается аналогично [4, Ch. 9, Theorem 9.2]. Теорема 1 доказана.

Литература

1. *Борисов А.* Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям I: характеристики точности // Информатика и её применения, 2019. Т. 13. Вып. 4. С. 68–75. doi: 10.14357/19922264190411.
2. *Борисов А.* Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям II: случай аддитивных шумов // Информатика и её применения, 2020. Т. 14. Вып. 1. С. 17–23. doi: 10.14357/19922264200103.
3. *Борисов А.* Численные схемы фильтрации марковских скачкообразных процессов по дискретизованным наблюдениям III: случай мультипликативных шумов // Информатика и её применения, 2020. Т. 14. Вып. 2. С. 10–18. doi: 10.14357/19922264200202.
4. *Liptser R., Shiryaev A.* Statistics of random processes I: General theory. — Berlin/Heidelberg: Springer, 2001. 427 p.
5. *Kallianpur G.* Stochastic filtering theory. — New York, NY, USA: Springer, 1980. 318 p.
6. *Yin G., Zhang Q., Liu Y.* Discrete-time approximation of Wonham filters // IET Control Theory A., 2004. No. 2. P. 1–10. doi: 10.1007/s11768-004-0017-7.
7. *Magnus J., Neudecker H.* Matrix differential calculus with applications in statistics and econometrics. — New York, NY, USA: Wiley, 2019. 504 p.
8. *Stoyanov J.* Counterexamples in probability. — Hoboken, NJ, USA: Wiley, 1997. 352 p.
9. *Borisov A., Sokolov I.* Optimal filtering of Markov jump processes given observations with state-dependent noises: Exact solution and stable numerical schemes // Mathematics, 2020. Vol. 8. Iss. 4. Art. No. 506. 22 p. doi: 10.3390/math8040506.
10. *Liptser R., Shiryaev A.* Theory of martingales. — Dortrecht: Springer, 1989. 792 p.
11. *Wong E., Hajek B.* Stochastic processes in engineering systems. — New York, NY, USA: Springer, 1985. 361 p.
12. *Elliott R., Moore J., Aggoun L.* Hidden Markov models: Estimation and control. — New York, NY, USA: Springer, 2010. 382 p.

Поступила в редакцию 05.03.2021

FILTERING OF MARKOV JUMP PROCESSES GIVEN COMPOSITE OBSERVATIONS I: EXACT SOLUTION

A. V. Borisov[1,2,3,4] and D. Kh. Kazanchyan[3]

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

[2]Moscow Aviation Institute (National Research University), 4 Volokolamskoe Shosse, Moscow 125080, Russian Federation

[3]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

[4]Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

Abstract: The first part of the series is devoted to the optimal filtering of the finite-state Markov jump processes (MJP) given the ensemble of the diffusion and counting observations. The noise intensity in the observable diffusion depends on the estimated MJP state. The special equivalent observation transformation converts them into the collection of the diffusion process of unit intensity, counting processes, and indirect measurements performed at some nonrandom discrete instants. The considered filtering estimate is expressed as a solution to the discrete-continuous stochastic differential system with the transformed observations on the right-hand side. The identifiability condition, under which MJP state can be reconstructed from indirect noisy observations precisely, is presented.

Keywords: Markov jump process; optimal filtering; multiplicative observation noises; stochastic differential equation; continuous and counting observations; identifiability condition

DOI: 10.14357/19922264210202

Acknowledgments

The work was supported in part by the Russian Foundation for Basic Research (project 19-07-00187 A). The research was conducted in accordance with the program of the Moscow Center for Fundamental and Applied Mathematics.

References

1. Borisov, A. 2019. Chislennyye skhemy fil'tratsii markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam I: kharakteristiki tochnosti [Numerical schemes of Markov jump process filtering given discretized observations I: Accuracy characteristics]. *Informatika i ee Primeneniya — Inform. Appl.* 13(4):68–75. doi: 10.14357/19922264190411.
2. Borisov, A. 2020. Chislennye skhemy fil'tratsii markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam II: sluchay additivnykh shuchmov [Numerical schemes of Markov jump process filtering given discretized observations II: Additive noises case]. *Informatika i ee primeneniya — Inform. Appl.* 14(1):17–23. doi: 10.14357/19922264200103.
3. Borisov, A. 2020. Chislennye skhemy fil'tratsii markovskikh skachkoobraznykh protsessov po diskretizovannym nablyudeniyam III: sluchay mul'tiplikativnykh shumov [Numerical schemes of Markov jump process filtering given discretized observations III: Multiplicative noises case]. *Informatika i ee primeneniya — Inform. Appl.* 14(2):10–18. doi: 10.14357/19922264200202.
4. Liptser, R., and A. Shiryaev. 2001. *Statistics of random processes I: General theory.* Berlin/Heidelberg: Springer. 427 p.
5. Kallianpur, G. 1980. *Stochastic filtering theory.* New York, NY: Springer. 318 p.
6. Yin, G., Q. Zhang, and Y. Liu. 2004. Discrete-time approximation of Wonham filters. *IET Control Theory A.* 2:1–10. doi: 10.1007/s11768-004-0017-7.
7. Magnus, J., and H. Neudecker. 2019. *Matrix differential calculus with applications in statistics and econometrics.* New York, NY: Wiley. 504 p.
8. Stoyanov, J. 1997. *Counterexamples in probability.* Hoboken, NJ: Wiley. 352 p.
9. Borisov, A., and I. Sokolov. 2020. Optimal filtering of Markov jump processes given observations with state-dependent noises: Exact solution and stable numerical schemes. *Mathematics* 8(4):506. 22 p. doi: 10.3390/math8040506.
10. Liptser, R., and A. Shiryaev. 1989. *Theory of martingales.* Dortrecht: Springer. 792 p.
11. Wong, E., and B. Hajek. 1985. *Stochastic processes in engineering systems.* New York, NY: Springer. 361 p.
12. Elliott, R., J. Moore, and L. Aggoun. 2010. *Hidden Markov models: Estimation and control.* New York, NY: Springer. 382 p.

Received March 5, 2021

Contributors

Borisov Andrey V. (b. 1965) — Doctor of Science in physics and mathematics, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Moscow Aviation Institute (National Research University), 4 Volokolamskoe Shosse, Moscow 125080, Russian Federation; professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; ABorisov@frccsc.ru

Kazanchyan Drastamat Kh. (b. 1994) — PhD student, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; drastamat94@gmail.com

ОБ ОДНОЙ НЕСТАЦИОНАРНОЙ МОДЕЛИ ОБСЛУЖИВАНИЯ С КАТАСТРОФАМИ И ТЯЖЕЛЫМИ ХВОСТАМИ*

А. И. Зейфман[1], Я. А. Сатин[2], И. А. Ковалёв[3]

Аннотация: Рассматривается нестационарная система массового обслуживания с катастрофами, одним сервером и специальными групповыми поступлениями требований, причем интенсивности увеличивающихся групп требований могут убывать достаточно медленно. Рассмотрен процесс $X(t)$, описывающий число требований в такой системе, доказано существование предельного режима распределения вероятностей состояний и предельного среднего для $X(t)$, получены оценки скорости сходимости к предельному режиму и предельному среднему. Получены оценки аппроксимации с помощью усечений конечными процессами. В качестве примера рассмотрена простая модель нестационарной системы с достаточно медленной скоростью убывания интенсивностей поступления групп требований, когда размер группы растет.

Ключевые слова: нестационарная система массового обслуживания; счетные марковские цепи; предельные характеристики; скорость сходимости; аппроксимации

DOI: 10.14357/19922264210203

1 Введение

Первоначальное описание исходной модели и первые исследования проведены в [1], соответствующая нестационарная ситуация была впервые изучена в [2]. Аналогичная модель в случае наличия катастрофических сбоев системы введена и изучена в [3]. При этом число требований в соответствующей системе обслуживания описывается процессом $X(t)$ (см. далее). В указанных работах изучается случай, когда интенсивности поступления групп требований экспоненциально убывают при увеличении размера группы. В настоящей работе исследуется ситуация «тяжелых хвостов», когда эти интенсивности убывают со степенной скоростью.

2 Описание модели

Подробное описание модели приведено в [3], здесь же отметим, что рассматриваемый процесс $X(t)$ является неоднородной марковской цепью с непрерывным временем, счетным пространством состояний $\{0, 1, 2, \dots\}$, а транспонированная матрица интенсивностей $A(t) = (a_{ij}(t))_{i,j=0}^{\infty}$ имеет вид:

$$A(t) = \begin{pmatrix} -\lambda(t) & \mu(t)+\gamma(t) & \gamma(t) & \gamma(t) & \dots \\ \lambda(t)b_1 & -(\lambda(t)B_2+\mu(t)+\gamma(t)) & \mu(t) & 0 & \dots \\ \lambda(t)b_2 & \lambda(t)b_2 & -(\lambda(t)B_3+\mu(t)+\gamma(t)) & \mu(t) & \dots \\ \lambda(t)b_3 & \lambda(t)b_3 & \lambda(t)b_3 & -(\lambda(t)B_4+\mu(t)+\gamma(t)) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Здесь $\lambda(t)$ — интенсивность поступления группы требований; $\mu(t)$ — интенсивность обслуживания одного требования; $\gamma(t)$ — интенсивность катастрофы (одномоментной потери всех требований в системе).

Далее, $\lambda(t)b_k$ — интенсивность поступления группы требований такой, что общее их число в системе оказывается равным k, т. е. если в системе уже есть $k-j$ требований, то это интенсивность одновременного поступления группы из j требований. При этом $B_k = \sum_{j=k}^{\infty} b_j$ при всех $k \geq 1$, а $B_1 = 1$. Все функции, определяющие интенсивности, предполагаются неотрицательными и локально ин-

*Исследование выполнено при финансовой поддержке Российского научного фонда (проект 19-11-00020).

[1]Вологодский государственный университет; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; Вологодский научный центр Российской академии наук, a_zeifman@mail.ru

[2]Вологодский государственный университет, yacovi@mail.ru

[3]Вологодский государственный университет, kovalev.iv96@yandex.ru

тегрируемыми на $[0,\infty)$. Кроме того, разумеется, предполагается выполненным условие

$$\sum_{k\geq 1} B_k = \sum_{k\geq 1} k b_k < \infty. \tag{1}$$

3 Получение оценок для процесса $X(t)$

Обозначив через $\mathbf{p}(t) = (p_0(t), p_1(t), \dots)^{\mathrm{T}}$ вектор вероятностей состояний для процесса $X(t)$, получаем прямую систему Колмогорова

$$\frac{d}{dt}\,\mathbf{p}(t) = A(t)\mathbf{p}(t),$$

которую в рассматриваемой ситуации удобно преобразовать к виду:

$$\frac{d}{dt}\,\mathbf{p}(t) = A^*(t)\mathbf{p}(t) + \mathbf{g}\,(t)\,,\ t \geq 0, \tag{2}$$

где $\mathbf{g}\,(t) = (\gamma\,(t)\,, 0, 0, \dots)^{\mathrm{T}}$, а $A^*(t)$ — матрица с элементами $a^*_{ij}(t)$,

$$a^*_{ij}\,(t) = \begin{cases} a_{0j}\,(t) - \gamma\,(t) & \text{при } i = 0\,; \\ a_{ij}\,(t) & \text{в остальных случаях.} \end{cases}$$

Далее будем предполагать, что найдется $\varepsilon > 0$ такое, что

$$\int_0^\infty (\gamma(t) - \varepsilon\lambda(t))\,dt = \infty. \tag{3}$$

Если, в частности, интенсивности постоянны, то (3) выполнено при положительном γ, а если 1-периодичны, то для выполнения (3) достаточно, чтобы $\int_0^1 \gamma(t)\,dt > 0$.

Как показано в [3], выполнение условия (3) гарантирует слабую эргодичность $X(t)$ в равномерной операторной топологии и оценку

$$\|\mathbf{p}^*\,(t) - \mathbf{p}^{**}\,(t)\| \leq e^{-\int_0^t \gamma(u)\,du}\,\|\mathbf{p}^*\,(0) - \mathbf{p}^{**}\,(0)\| \leq \\ \leq 2e^{-\int_0^t \gamma(u)\,du},\ t \geq 0, \tag{4}$$

справедливую при любых начальных условиях $\mathbf{p}^*\,(0)$ и $\mathbf{p}^{**}\,(0)$.

Однако, как и в предыдущих работах, интерес представляет не само наличие предельного режима, а возможность его построения.

Для получения нужных свойств и оценок потребуются некоторые вспомогательные «взвешенные» нормы.

Положим $d_0 = 1$, и пусть $\{d_k\}$ — неубывающая последовательность, $k \geq 0$. Рассмотрим диагональную матрицу $\Lambda = \operatorname{diag}(d_0, d_1, d_2, \dots)$.

Тогда из (2) получим уравнение:

$$\frac{d}{dt}\,\tilde{\mathbf{p}}(t) = \tilde{A}^*(t)\tilde{\mathbf{p}}(t) + \tilde{\mathbf{g}}\,(t)\,, \tag{5}$$

где $\tilde{\mathbf{p}}(t) = \Lambda\mathbf{p}(t)$; $\tilde{A}(t) = \Lambda A(t)\Lambda^{-1}$; $\tilde{\mathbf{g}}(t) = \Lambda\mathbf{g}(t)$.

Далее будем оценивать логарифмическую норму оператора $\tilde{A}(t)$. Если обозначить через $-\tilde{\alpha}_k(t)$ сумму всех элементов k-го столбца матрицы $\tilde{A}(t)$, то получим

$$\tilde{\alpha}_0(t) \geq \gamma(t) - \lambda(t)\sum_{j=1}^{\infty}\left(\frac{d_j}{d_0} - 1\right) := \beta(t);$$

$$\tilde{\alpha}_k(t) \geq \gamma(t) - \lambda(t)\sum_{j=k+1}^{\infty}\left(\frac{d_j}{d_k} - 1\right) \geq \tilde{\alpha}_0(t) = \beta(t), \\ k \geq 1\,.$$

Из условия (1) вытекает, что для любого $\varepsilon > 0$ найдется натуральное N такое, что

$$\sum_{k\geq N}(k-1)\,b_k < \varepsilon.$$

Положим теперь $d_k = 1$, если $k < N$, и $d_k = k$ при $k \geq N$.

Тогда логарифмическая норма оператора $\tilde{A}(t)$ равна

$$-\beta^*(t) = \sup_i\left\{\tilde{a}_{ii}(t) + \sum_{j\neq i}\tilde{a}_{ji}(t)\right\} = -\beta(t) \leq \\ \leq -(\gamma(t) - \varepsilon\lambda(t))\,.$$

Следовательно, вместо (4) получаем

$$\|\tilde{\mathbf{p}}^*\,(t) - \tilde{\mathbf{p}}^{**}\,(t)\| \leq e^{-\int_0^t \beta(u)\,du}\,\|\tilde{\mathbf{p}}^*\,(0) - \tilde{\mathbf{p}}^{**}\,(0)\|\,, \\ t \geq 0\,.$$

Далее, сравнивая соответствующие нормы и математические ожидания, получаем такое утверждение.

Теорема 1. *Пусть выполнены условия* (1) *и* (3). *Тогда $X(t)$ слабо эргодичен, имеет предельное среднее и справедливы следующие оценки скорости сходимости*:

$$\|\mathbf{p}^*\,(t) - \mathbf{p}^{**}\,(t)\| \leq e^{-\int_0^t \beta(u)\,du}\,\|\tilde{\mathbf{p}}^*\,(0) - \tilde{\mathbf{p}}^{**}\,(0)\|\,, \\ t \geq 0\,;$$

$$|E(t,k) - E(t,0)| \leq kNe^{-\int_0^t \beta(u)\,du}, \qquad t \geq 0\,,$$

где $E(t,j)$ — математическое ожидание (среднее число требований) для $X(t)$ при условии, что $X(0)=j$.

Оценим теперь само предельное среднее. Дополнительно предположим выполнение условий

$$e^{\int_s^t \beta(u)\,du} \le Re^{-a(t-s)},\ \ \gamma(t)\le\theta, \tag{6}$$

для всех $0\le s\le t$ при некоторых положительных R, a и θ.

Обозначим через $\tilde U(t,s)$ оператор Коши уравнения (5). Тогда получим

$$\tilde{\mathbf{p}}(t)=\tilde U(t,0)\tilde{\mathbf{p}}(0)+\int\limits_0^t \tilde U(t,\tau)\tilde{\mathbf{g}}\,(\tau)\,d\tau,$$

откуда

$$\limsup_{t\to\infty}\tilde{\mathbf{p}}(t)\le\limsup_{t\to\infty}\int\limits_0^t Re^{-a(t-\tau)}\theta\,d\tau\le\frac{R\theta}{a}\,.$$

Тогда имеем

Следствие 1. Пусть выполнены условия (1) и (6). Тогда при любом k справедлива следующая оценка:

$$\limsup_{t\to\infty}E(t,k)\le\frac{NR\theta}{a}\,.$$

4 Аппроксимация усечениями

В заключение рассмотрим вопрос о построении предельного режима и предельного среднего с помощью аппроксимации усеченными процессами. Получение не зависящих от времени оценок при таких аппроксимациях описано в 4, 5].

Аналогично этим работам, будем отождествлять конечные векторы и счетные векторы с теми же ненулевыми координатами. Рассмотрим «усеченную» матрицу (для краткости зависимость от t не записываем)

$$A_K^*=\begin{pmatrix} -\lambda B_1^*-\gamma & \mu & 0 & 0 & \dots\\ \lambda b_1 & -(\lambda B_2^*+\mu+\gamma) & \mu & 0 & \dots\\ \dots & \dots & \dots\dots\dots & & \\ \lambda b_K & \lambda b_K & \dots\dots\dots & & \end{pmatrix}$$
$$\begin{pmatrix} \dots & 0 & & \\ \dots & 0 & & \\ \dots\dots\dots & \dots & \dots\\ \dots\dots\dots & -(\lambda B_K^*+\mu+\gamma) & \end{pmatrix},$$

где $B_k^*=\sum_{j=k}^K b_j$.

Запишем аналогичную (2) систему для усеченного процесса в виде:

$$\frac{d}{dt}\,\mathbf{p}_K(t)=A^*(t)\mathbf{p}_K(t)+\mathbf{g}\,(t)+ +(A_K^*(t)-A^*(t))\,\mathbf{p}_K(t),\ \ t\ge 0\,.$$

Тогда

$$\mathbf{p}_K(t)=U(t,0)\mathbf{p}(0)+\int\limits_0^t U(t,\tau)\mathbf{g}\,(\tau)\,d\tau+ +\int\limits_0^t U(t,\tau)\,(A_K^*(\tau)-A^*(\tau))\,\mathbf{p}_K(\tau)\,d\tau= =\mathbf{p}(t)+\int\limits_0^t U(t,\tau)\,(A_K^*(\tau)-A^*(\tau))\,\mathbf{p}_K(\tau)\,d\tau,$$

где $U(t,s)$ — оператор Коши уравнения (2).

Следовательно, в любой норме справедлива оценка:

$$\|\mathbf{p}\,(t)-\mathbf{p}_K\,(t)\|\le \le\int\limits_0^t\|U(t,\tau)\|\,\|\,(A_K^*(\tau)-A^*(\tau))\,\mathbf{p}_K(\tau)\|\,d\tau. \tag{7}$$

Рассмотрим норму $\|\mathbf{x}\|_\Lambda=\|\Lambda\mathbf{x}\|$, тогда $\|\tilde U(t,s)\|=\|U(t,s)\|_\Lambda\le Re^{-a(t-s)}$.

Для оценки второго множителя под знаком интеграла в (7) отметим, что в левом верхнем квадрате матрицы $A_K^*-A^*$ (где оба индекса не превосходят K) ненулевыми являются только диагональные элементы, каждый из которых равен $-\lambda B_K$. Значит,

$$(A_K^*(\tau)-A^*(\tau))\,\mathbf{p}_K(\tau)= =-\lambda(\tau)B_K\,(p_0(\tau),\dots,p_K(\tau))^{\mathrm T}\,.$$

А тогда, предполагая, что $\lambda(t)\le\theta$ при всех t, получаем

$$\|\,(A_K^*(\tau)-A^*(\tau))\,\mathbf{p}_K(\tau)\|\le \le B_K\theta\sum_{k\le K}d_kp_k(\tau)\le B_KN\theta\,.$$

Тогда правая часть в (7) в Λ-норме не превосходит $B_KNR\theta/a$ и получаем следующее утверждение.

Теорема 2. *Пусть выполнены условия* (1) *и* (6). *Тогда при $X(0)=0$ справедливы следующие оценки*:

$$\|\mathbf{p}\,(t)-\mathbf{p}_K\,(t)\|\le\frac{B_KNR\theta}{a}\to 0\text{ при }k\to\infty\,;$$

$$|E(t,0)-E_K(t,0)|\le\frac{B_KN^2R\theta}{a}\to 0\text{ при }k\to\infty.$$

5 Численный пример

Рассмотрим описанную модель с интенсивностями $\gamma=1$, $\lambda(t)=1+\sin 2\pi t$ и $\mu(t)=1+ +\cos 2\pi t$, предполагая при этом, что $b_k=4/((k(k+ +1)(k+2)))$, т. е. убывание имеет степенной характер.

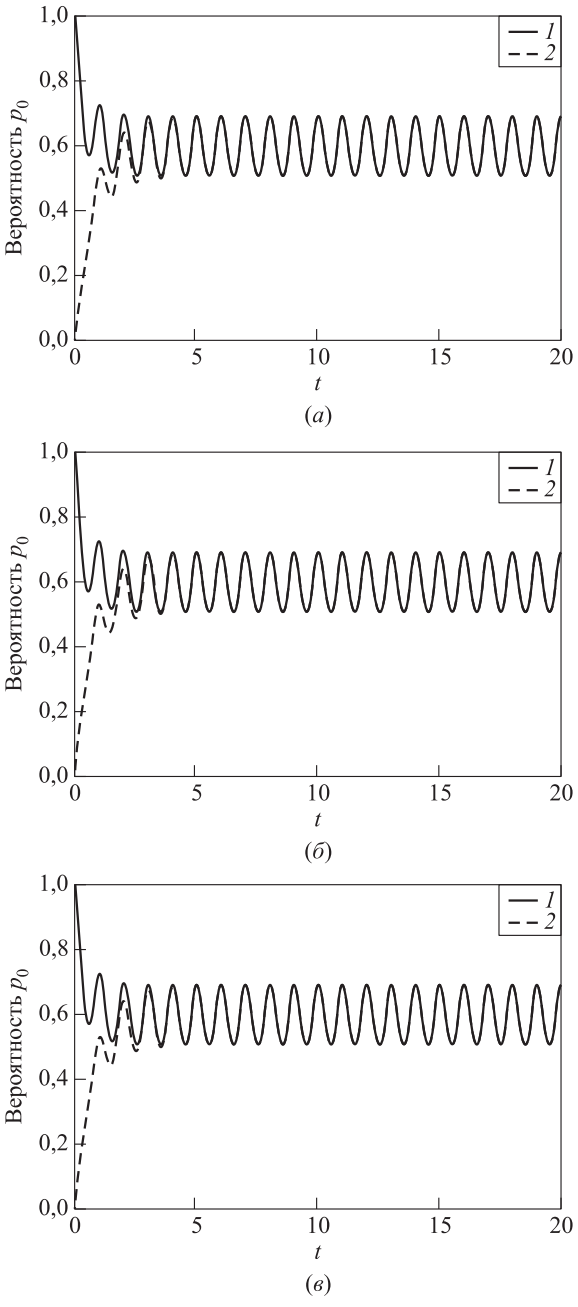


Рис. 1 Поведение вероятности $p_0(t)$ для усеченного процесса, интервал $[0, 20]$: (*а*) 100 состояний; (*б*) 200; (*в*) 300 состояний; *1* — $x(0) = 0$; *2* — $x(0) = 100$

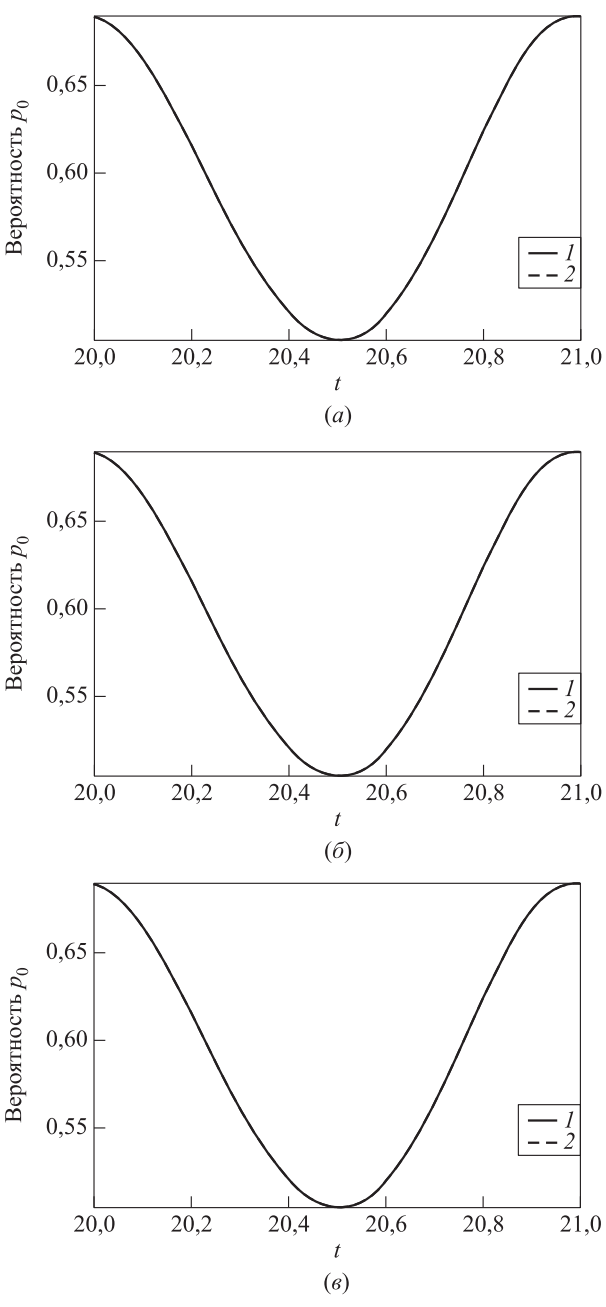


Рис. 2 Предельная вероятность $p_0(t)$ для усеченного процесса, интервал $[20, 21]$: (*а*) 100 состояний; (*б*) 200; (*в*) 300 состояний; *1* — $x(0) = 0$; *2* — $x(0) = 100$

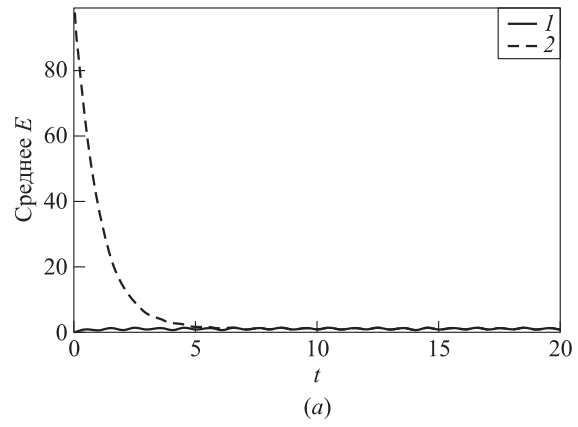
На рис. 1–4 показано поведение вероятности отсутствия требований в системе $p_0(t)$ и среднего числа требований в системе $E(t, k)$ для усеченных процессов с числом состояний 100, 200 и 300.

Можно отметить, что погрешность вектора вероятностей состояний при усечениях, соответствующих $K = 100$ и 200, получается $1{,}2 \cdot 10^{-2}$ и $1{,}2 \cdot 10^{-3}$ соответственно, а для средних — $7{,}2 \cdot 10^{-2}$ и $7{,}2 \cdot 10^{-3}$ соответственно.

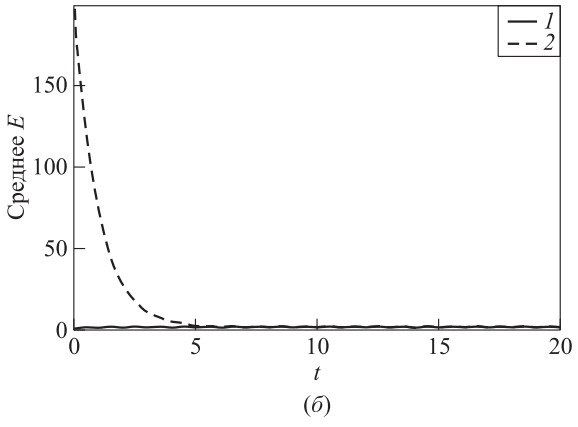
Литература

1. *Marin A., Rossi S.* A queueing model that works only on the biggest jobs // 16th European Computer Performance Engineering Workshop Revised Selected Papers / Eds. M. Gribaudo, M. Iacono, T. Phung-Duc, R. Razumchik. — Lecture notes in computer science ser. — Springer, 2020. Vol. 12039. P. 118–132.

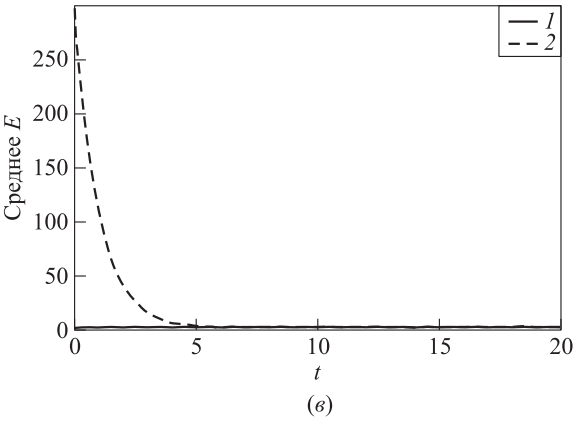
2. *Zeifman A. I., Razumchik R. V., Satin Y. A., Kovalev I. A.* Ergodicity bounds for the Markovian queue with time-



(*а*)

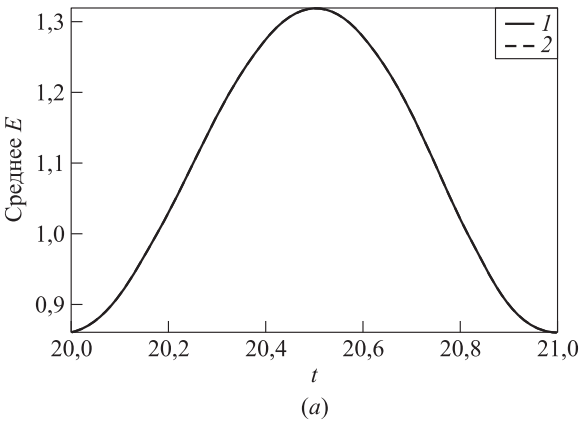


(*б*)

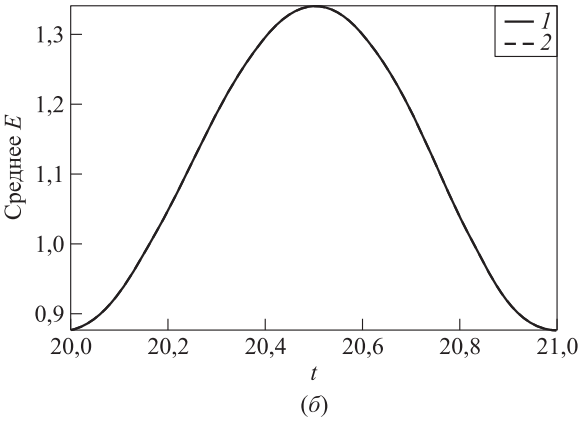


(*в*)

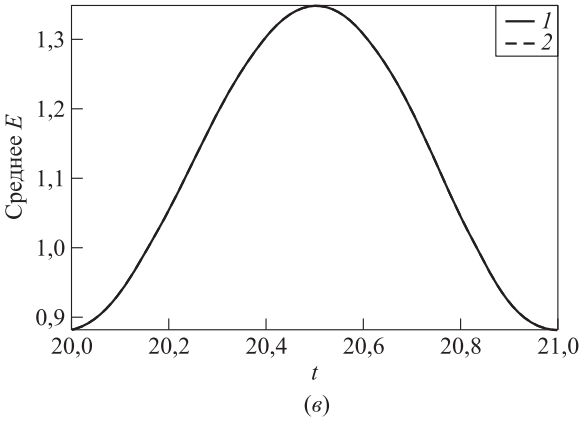
Рис. 3 Поведение среднего $E(t,k)$ для усеченного процесса, интервал $[0,20]$: (*а*) 100 состояний; (*б*) 200; (*в*) 300 состояний; *1* — $x(0)=0$; *2* — $x(0)=100$



(*а*)



(*б*)



(*в*)

Рис. 4 Предельное среднее $E(t,k)$ для усеченного процесса, интервал $[20,21]$: (*а*) 100 состояний; (*б*) 200; (*в*) 300 состояний; *1* — $x(0)=0$; *2* — $x(0)=100$

varying transition intensities, batch arrivals and one queue skipping policy // Appl. Math. Comput., 2021. Vol. 395. Art. 125846.

3. *Zeifman A., Satin Y., Kovalev I., Razumchik R., Korolev V.* Facilitating numerical solutions of inhomogeneous continuous time Markov chains using ergodicity bounds obtained with logarithmic norm method // Mathematics, 2021. Vol. 9. Iss. 1. Art. 42. 20 p.

4. *Zeifman A., Satin Y., Korolev V., Shorgin S.* On truncations for weakly ergodic inhomogeneous birth and death processes // Int. J. Appl. Math. Comp., 2014. Vol. 24. No. 3. P. 503–518.

5. *Зейфман А. И., Коротышева А. В., Королев В. Ю., Сатин Я. А.* Оценки погрешности аппроксимаций неоднородных марковских цепей с непрерывным временем // Теория вероятностей и ее применения, 2016. Т. 61. № 3. С. 563–569.

Поступила в редакцию 07.03.2021

ON ONE NONSTATIONARY SERVICE MODEL WITH CATASTROPHES AND HEAVY TAILS

A. I. Zeifman[1,2,3], Ya. A. Satin[1], and I. A. Kovalev[1]

[1]Department of Applied Mathematics, Vologda State University, 15 Lenin Str., Vologda 160000, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation
[3]Vologda Research Center of the Russian Academy of Sciences, 56A Gorky Str., Vologda 160014, Russian Federation

Abstract: The paper considers the nonstationary queuing system with catastrophes, one server, and special group arrivals of requests. The intensities of increasing groups of requests can decrease rather slowly. The process $X(t)$, which describes the number of requirements in such system, is considered, the existence of a limiting regime of the probability distribution of states and a limiting average for $X(t)$ is proved, and estimates of the rate of convergence to the limiting regime and the limiting average are obtained. Approximation estimates are obtained using truncations by finite processes. As an example, the authors consider a simple model of a nonstationary system with a rather slow rate of decrease in the arrival rates of customer groups when the group size grows.

Keywords: nonstationary queuing system; countable Markov chains; limiting characteristics; rate of convergence; approximation

DOI: 10.14357/19922264210203

Acknowledgments

This work was financially supported by the Russian Science Foundation (grant No. 19-11-00020).

References

1. Marin, A., and S. Rossi. 2020. A queueing model that works only on the biggest jobs. *16th European Computer Performance Engineering Workshop Revised Selected Papers*. Eds. M. Gribaudo, M. Iacono, T. Phung-Duc, and R. Razumchik. Lecture notes in computer science ser. Springer. 12039:118–132.
2. Zeifman, A. I., R. V. Razumchik, Y. A. Satin, and I. A. Kovalev. 2021. Ergodicity bounds for the Markovian queue with time-varying transition intensities, batch arrivals and one queue skipping policy. *Appl. Math. Comput.* 395:125846. 11 p.
3. Zeifman, A., Y. Satin, I. Kovalev, R. Razumchik, and V. Korolev. 2021. Facilitating numerical solutions of inhomogeneous continuous time Markov chains using ergodicity bounds obtained with logarithmic norm method. *Mathematics* 9(1):42. 20 p.
4. Zeifman, A., Y. Satin, V. Korolev, and S. Shorgin. 2014. On truncations for weakly ergodic inhomogeneous birth and death processes. *Int. J. Appl. Math. Comp.* 24(3):503–518.
5. Zeifman, A. I., A. V. Korotysheva, V. Y. Korolev, and Ya. A. Satin. 2017. Truncation bounds for approximations of inhomogeneous continuous-time Markov chains. *Theor. Probab. Appl.* 61(3):513–520.

Received March 7, 2021

Contributors

Zeifman Alexander I. (b. 1954) — Doctor of Science in physics and mathematics, professor, Head of Department, Vologda State University, 15 Lenin Str., Vologda 160000, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; principal scientist, Vologda Research Center of the Russian Academy of Sciences, 56A Gorky Str., Vologda 160014, Russian Federation; a_zeifman@mail.ru

Satin Yacov A. (b. 1978) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Applied Mathematics, Vologda State University, 15 Lenin Str., Vologda 160000, Russian Federation; yacovi@mail.ru

Kovalev Ivan A. (b. 1996) — PhD student, Department of Applied Mathematics, Vologda State University, 15 Lenin Str., Vologda 160000, Russian Federation; kovalev.iv96@yandex.ru

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2021. Т. 15. Вып. 2. С. 26–29

МНОГОМЕРНЫЕ РАСПРЕДЕЛЕНИЯ ВЫХОДЯЩИХ ПОТОКОВ В СИСТЕМЕ С АБСОЛЮТНЫМ ПРИОРИТЕТОМ*

В. Г. Ушаков[1], Н. Г. Ушаков[2]

Аннотация: Изучена однолинейная система массового обслуживания с бесконечным числом мест для ожидания, произвольным распределением времени обслуживания и пуассоновскими входящими потоками требований. Между требованиями разных потоков действует дисциплина абсолютного приоритета с обслуживанием заново прерванного требования. Методом вложенных цепей Маркова исследуется многомерный случайный процесс, компоненты которого — число требований фиксированного приоритета в системе и длительность интервала времени между последовательными моментами ухода из системы требований этого приоритета. Найдены конечномерные распределения указанного процесса. В качестве следствия получены преобразования Лапласа–Стилтьеса одномерных и двумерных распределений выходящего потока требований каждого приоритета в стационарном режиме.

Ключевые слова: выходящий поток; абсолютный приоритет; вложенная цепь Маркова; одноканальная система

DOI: 10.14357/19922264210204

1 Введение

Свойства выходящих из системы обслуживания потоков требований важны при решении многих задач, среди которых исследование сетей массового обслуживания, статистический анализ параметров системы, оптимальное управление работой системы.

Вероятностные свойства выходящих потоков в приоритетных системах изучены пока недостаточно глубоко. В частности, практически нет работ, посвященных исследованию многомерных распределений. Методы, применяемые в работах [1–3] при нахождении одномерных распределений, обобщить на случай многомерных распределений не удается. В работе [4] предложен метод нахождения любых конечномерных распределений в системе с относительным приоритетом. В настоящей работе этот метод применен для анализа системы с абсолютным приоритетом и обслуживанием прерванного требования заново.

2 Обозначения и определения

Рассматривается система обслуживания типа $M_r|G_r|1|\infty$ с абсолютным приоритетом и обслуживанием прерванного требования заново. Считаем, что потоки пронумерованы в порядке убывания их важности, т. е. требования из потока с меньшим номером обладают более высоким приоритетом. Пусть $a_1,\dots,a_r$ — интенсивности, а $B_1(x),\dots,B_r(x)$ — функции распределения времен обслуживания требований потоков $1,\dots,r$ соответственно. Обозначим

$$\beta_i(s)=\int\limits_0^\infty e^{-sx}dB(x);\quad \beta_{ij}=\int\limits_0^\infty x^j dB_i(x);$$

$$\sigma_k=a_1+\dots+a_k,\quad k=1,\dots,r;$$

$$\sigma_0=0;\quad \sigma=\sigma_r;\quad \rho_{11}=a_1\beta_{11};$$

$$\rho_{i1}=a_1\beta_{11}+a_2\sigma_1^{-1}\left(\beta_2^{-1}(\sigma_1)-1\right)+\dots$$
$$\dots+a_i\sigma_{i-1}^{-1}\left(\beta_i^{-1}(\sigma_{i-1})-1\right),\quad i=2,\dots,r.$$

Пусть, далее, t_{iN} — момент ухода из системы N-го требования приоритета i (нумерация требований проводится для каждого приоритета отдельно в порядке их ухода из системы), $t_{i0}=0$, $\tau_{iN}=t_{iN}-t_{i,N-1}$, $L_i(t)$ — число требований i-го потока (приоритета i) в системе в момент времени t, $i=1,2,\dots r$, $N=1,2,\dots$

Всюду в дальнейшем будем считать выполненным условие эргодичности $\rho_{r1}<1$. Положим

*Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (проект 075-15-2019-1621).

[1]Факультет вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; Федеральный исследовательский центр «Информатика и управление» Российской академии наук, vgushakov@mail.ru

[2]Институт проблем технологии микроэлектроники и особочистых материалов Российской академии наук; Норвежский научно-технологический университет, Тронхейм, Норвегия, ushakov@math.ntnu.no

$$P_i(n,x) = \lim_{N\to\infty} \mathbb{P}\left(L_i(t_{iN}+0) = n, \tau_{iN} < x\right);$$

$$Q_i(n,m,x,y) = \lim_{N\to\infty} \mathbb{P}\left(L_i(t_{iN}+0) = n, L_i(t_{i,N-1}+0) = m, \tau_{iN} < x, \tau_{i,N-1} < y\right);$$

$$p_i(z,s) = \int_0^\infty e^{-sx} \sum_{n=0}^\infty z^n d_x P_i(n,x);$$

$$q_i(w,z,s_1,s_2) = \int_0^\infty\int_0^\infty e^{-s_1x}e^{-s_2y} \sum_{n=0}^\infty\sum_{m=0}^\infty w^n z^m d_x d_y Q_i(n,m,x,y);$$

$$f_i(s) = \lim_{N\to\infty} \int_0^\infty e^{-sx} d\mathbb{P}(\tau_{iN} < x);$$

$$g_i(s_1,s_2) = \lim_{N\to\infty} \int_0^\infty\int_0^\infty e^{-s_1x}e^{-s_2y} d_x d_y \mathbb{P}(\tau_{iN} < x, \tau_{i,N-1} < y).$$

3 Предварительные результаты

При изучении выходящих потоков будут использованы некоторые известные результаты (см., например, [5]) для рассматриваемой системы массового обслуживания. Пусть Π_k, H_k и Π_{ki} — k-период, k-цикл и ki-период соответственно:

$$\pi_k(s) = \mathbb{E}e^{-s\Pi_k};\ h_k(s) = \mathbb{E}e^{-sH_k};\ \pi_{ki}(s) = \mathbb{E}e^{-s\Pi_{ki}}.$$

Тогда

$$\pi_k(s) = \sum_{i=1}^k a_i \sigma_k^{-1} \pi_{ki}(s);$$

$$h_k(s) = \beta_k(s+\sigma_{k-1}) \times \left(1 - \frac{\sigma_{k-1}\pi_{k-1}(s)}{s+\sigma_{k-1}}\left(1-\beta_k(s+\sigma_{k-1})\right)\right)^{-1},$$

а $\pi_{ki}(s)$ — единственное решение системы уравнений

$$\pi_k(s) = \beta_i\left(s+\sigma_k - \sum_{j=i}^k a_j\pi_{kj}(s)\right) - \sum_{j=1}^{i-1}\pi_{kj}(s) \times \pi_{ki}(s)\frac{1-\beta_i\left(s+\sigma_k-\sum_{j=i}^k a_j\pi_{kj}(s)\right)}{\left(s+\sigma_k-\sum_{j=i}^k a_j\pi_{kj}(s)\right)}, \quad i = 1,\ldots,k.$$

Пусть, далее, $W_i(t)$ — виртуальное время ожидания для требований приоритета i в момент времени t:

$$\omega_i^*(s,t) = \mathbb{E}e^{-sW_i(t)};$$

$$\omega_i(s,q) = \int_0^\infty e^{-qt}\omega_i^*(s,t)\,dt.$$

Тогда

$$\omega_1(s,q) = \frac{1-sp_0^{(1)}(q)}{q-\alpha_1(s)},$$

$$\omega_i(s,q) = \frac{1-sp_0^{(i)}(q)-(1-\pi_{i-1}(s))p_1^{(i)}(q)}{q-\alpha_i(s)}, \quad i = 2,\ldots,r.$$

Здесь

$$\alpha_k(s) = s - a_k + a_k h_k(s);$$

$$p_0^{(k)}(q) = (q+\sigma_k-\sigma_k\pi_k(q))^{-1},\ \ k = 1,\ldots,r;$$

$$p_1^{(k)}(q) = \frac{1-(q+a_k-a_k\pi_{kk}(q))p_0^{(k)}(q)}{1-\pi_{k-1}(q+a_k-a_k\pi_{kk}(q))}, \quad k = 2,\ldots,r.$$

4 Основные результаты

В дальнейшем будут изучены выходящие потоки требований приоритетов $2,\ldots,r$. Для требований первого приоритета справедливы результаты для неприоритетной системы, в которую поступает только первый поток. Основные результаты работы содержатся в приводимых ниже двух теоремах.

Теорема 1. *Функции $p_i(z,s)$ и $q_i(z,w,s_1,s_2)$ определяются соотношениями*:

$$p_i(z,s) = \left(z^{-1}(p_i(z,0)-p_i(0,0)) + a_i p_i(0,0)\right) \times \omega_{i-1}\left(s+a_{i-1}-a_{i-1}\pi_{i-1,i-1}(s+a_i-a_iz)+ a_i-a_iz, s+a_i\right)h_i(s+a_i-a_iz); \quad (1)$$

$$q_i(z,w,s_1,s_2) = \left(z^{-1}(p_i(zw,s_2)-p_i(0,s_2)) + a_i p_i(0,s_2)\right) \times \omega_{i-1}\left(s_1+a_{i-1}-a_{i-1}\pi_{i-1,i-1}(s_1+a_i-a_iz)+ a_i-a_iz, s_1+a_i\right)h_i(s_1+a_i-a_iz). \quad (2)$$

Функция $p_i(z,0)$ равна

$$p_i(z,0) = \frac{p_i(0,0)h_i(a_i-a_iz)}{z-h_i(a_i-a_iz)}\left(a_iz\omega_{i-1}(a_{i-1}- - a_{i-1}\pi_{i-1,i-1}(a_i-a_iz)+a_i-a_iz, a_i)-1\right), \quad (3)$$

где

$$p_2(0,0) = a_2^{-1}(1-\rho_{21})(\sigma_2-a_1\pi_1(a_2));$$

$$p_i(0,0) = a_i^{-1}(1-\rho_{i1})(\sigma_i-\sigma_{i-1}\pi_{i-1}(a_i)) \times \left(1-\rho_{i-2,1}+\sigma_{i-2}^{-1}\rho_{i-2,1}\times\right.$$

$$\times \frac{\sigma_{i-1} - \sigma_{i-1}\pi_{i-1}(a_i) - a_{i-1} + a_{i-1}\pi_{i-1,i-1}(a_i)}{1 - \pi_{i-2}(a_i + a_{i-1} - a_{i-1}\pi_{i-1,i-1}(a_i))}\Bigg)^{-1},$$

$$i = 3, \ldots, r;$$

$$p_i(0,s) = (v_i + a_i p_i(0,0)\omega_{i-1} \times$$

$$\times (s + a_{i-1} - a_{i-1}\pi_{i-1,i-1}(s + a_i) + a_i, s + a_i)) \times$$

$$\times h_i(s + a_i);$$

$$v_i = p_i(0,0)\left(h_i^{-1}(a_i) - a_i \times\right.$$

$$\left.\times \omega_{i-1}(a_{i-1} - a_{i-1}\pi_{i-1,i-1}(a_i) + a_i, a_i)\right).$$

Доказательство. Рассматривая два соседних момента ухода требований i-го приоритета из системы, имеем

$$P_i(n,x) = \sum_{j=1}^{n+1} P_i(j,\infty) \int_0^x e^{-a_i u} \frac{(a_i u)^{n-j+1}}{(n-j+1)!}\, dH_i(u) +$$

$$+ P_i(0,\infty) \int_0^x G_i(u,n,x-u)\, d\left(1 - e^{-a_i u}\right), \quad (4)$$

где $G_i(u,k,v)$ — вероятность того, что первое требование приоритета i уйдет из системы через время, меньшее чем $u+v$, и в момент его ухода будет k требований этого приоритета, при условии что оно поступает в момент u, а в начальный момент система свободна.

Положим

$$g_i(u,z,s) = \sum_{k=0}^{\infty} z^k \int_0^{\infty} e^{-sx} d_x G_i(u,k,x).$$

Тогда при $i \geqslant 2$ имеем

$$g_i(u,z,s) = \omega_{i-1}^*(s + a_{i-1} - a_{i-1}\pi_{i-1,i-1}(s + a_i - a_i z) + a_i - a_i z, u)\, h_i(s + a_i - a_i z).$$

Переходя в (4) к производящим функциям и преобразованиям Лапласа–Стилтьеса, получаем (1). Из (1) при $s = 0$ получаем (3). Устремляя в (3) z к единице, находим $p_i(0,0)$.

Рассмотрим теперь три последовательных момента ухода из системы требований i-го приоритета. Имеем

$$Q_i(n_1,n_2,x,y) = P_i(n_2,y) \int_0^x e^{-a_i u} \frac{(a_i u)^{n_1+1-n_2}}{(n_1+1-n_2)!}\, dH_i(u),$$

$$n_2 \geqslant 1, \quad n_1 \geqslant n_2 - 1; \quad (5)$$

$$Q_i(n_1,0,x,y) = Q_i(0,y) \int_0^x G_i(u,n_1,x-u)\, d\left(1 - e^{-a_i u}\right). \quad (6)$$

Переходя в (5) и (6) к производящим функциям и преобразованиям Лапласа–Стилтьеса, получаем (2).

Теорема 2. *Справедливы следующие соотношения*:

$$f_i(s) = (1 - p_i(0,0) + a_i p_i(0,0) \times$$

$$\times \omega_{i-1}(s + a_{i-1} - a_{i-1}\pi_{i-1,i-1}(s), s + a_i))\, h_i(s),$$

$$g_i(s_1,s_2) = (1 - p_i(0,s_2) + a_i p_i(0,s_2) \times$$

$$\times \omega_{i-1}(s_1 + a_{i-1} - a_{i-1}\pi_{i-1,i-1}(s_1), s_1 + a_i)) \times$$

$$\times h_i(s_1).$$

Доказательство непосредственно вытекает из результатов теоремы 1 и соотношений $f_i(s) = p_i(1,s)$ и $g_i(s_1,s_2) = q_i(1,1,s_1,s_2)$.

Литература

1. *Nain P.* Interdeparture times from a queuing system with preemptive resume priority // Perform. Evaluation, 1984. Vol. 4. Iss. 2. P. 93–98. doi: 10.1016/0166-5316(84)90003-8.
2. *Stanford D. A.* Interdeparture time distributions in the non-preemptive priority $\Sigma M_i|G_i|1$ queue // Perform. Evaluation, 1991. Vol. 12. Iss. 2. P. 43–60.
3. *Stanford D. A.* Waiting and interdeparture times in priority queues with Poisson- and general-arrival streams // Oper. Res., 1995. Vol. 45. Iss. 5. P. 725–735.
4. *Ушаков В. Г., Ушаков Н. Г.* Выходящие потоки в однолинейной системе с относительным приоритетом // Информатика и её применения, 2019. Т. 13. Вып. 4. С. 42–47. doi: 10.14357/19922264190407.
5. *Матвеев В. Ф., Ушаков В. Г.* Системы массового обслуживания. — М.: Изд-во Московского ун-та, 1984. 240 с.

Поступила в редакцию 07.04.2021

THE MULTIVARIATE DISTRIBUTIONS OF OUTPUT STREAMS IN A QUEUEING SYSTEM WITH PREEMPTIVE REPEAT PRIORITY

V. G. Ushakov[1,2] and N. G. Ushakov[3,4]

[1]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation
[3]Institute of Microelectronics Technology and High-Purity Materials of the Russian Academy of Sciences, 6 Academician Osipyan Str., Chernogolovka, Moscow Region 142432, Russian Federation
[4]Norwegian University of Science and Technology, 15A S. P. Andersensvei, Trondheim 7491, Norway

Abstract: The paper studies a single server queuing system with r types of customers, preemptive repeat priority, and an infinite number of positions in the queue. The arrival stream of customers of each type is a Poisson stream. Each type has its own generally distributed service time characteristics. The main result is the Laplace–Stieltjes transform of one- and two-dimensional stationary distribution functions of the interdeparture times for each type of customers. The analysis of the output process is carried out by the method of embedded Markov chains. As embedded times, successive moments of the end of service of the same type customers are selected. From a practical perspective, an accurate characterization of the interdeparture time process is necessary when studying open networks of queues.

Keywords: output stream; preemptive repeat priority; embedded Markov chain; single server

DOI: 10.14357/19922264210204

Acknowledgments

The research was supported by the Ministry of Science and Higher Education of the Russian Federation, project No. 075-15-2019-1621.

References

1. Nain, P. 1984. Interdeparture times from a queuing system with preemptive resume priority. *Perform. Evaluation* 4(2):93–98. doi: 10.1016/0166-5316(84)90003-8.
2. Stanford, D. A. 1991. Interdeparture time distributions in the non-preemptive priority $\Sigma\ M_i|G_i|1$ queue. *Perform. Evaluation* 12(2):43–60.
3. Stanford, D. A. 1995. Waiting and interdeparture times in priority queues with Poisson- and general-arrival streams. *Oper. Res.* 45(5):725–735.
4. Ushakov, V. G., and N. G. Ushakov. 2019. Vykhodyashchie potoki v odnolineynoy sisteme s otnositel'nym prioritetom [The output streams in the single server queueing system with a head of the line priority]. *Informatika i ee Primeneniya — Inform. Appl.* 13(4):42–47. doi: 10.14357/19922264190407.
5. Matveev, V. F., and V. G. Ushakov. 1984. *Sistemy massovogo obsluzhivaniya* [Queueing systems]. Moscow: MSU Publs. 240 p.

Received April 7, 2021

Contributors

Ushakov Vladimir G. (b. 1952) — Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; vgushakov@mail.ru

Ushakov Nikolai G. (b. 1952) — Doctor of Science in physics and mathematics, leading scientist, Institute of Microelectronics Technology and High-Purity Materials of the Russian Academy of Sciences, 6 Academician Osipyan Str., Chernogolovka, Moscow Region 142432, Russian Federation; professor, Norwegian University of Science and Technology, 15A S. P. Andersensvei, Trondheim 7491, Norway; ushakov@math.ntnu.no

АНАЛИЗ НЕСМЕЩЕННОЙ ОЦЕНКИ СРЕДНЕКВАДРАТИЧНОГО РИСКА МЕТОДА БЛОЧНОЙ ПОРОГОВОЙ ОБРАБОТКИ*

О. В. Шестаков[1]

Аннотация: Методы обработки сигналов и изображений, основанные на вейвлет-разложении и пороговой обработке, приобрели большую популярность при решении задач подавления шума и компрессии. Это объясняется их адаптивностью к локальным особенностям исследуемых функций, высокой скоростью алгоритмов обработки и оптимальностью получаемых оценок. В данной работе рассмотрен метод блочной пороговой обработки, в котором коэффициенты разложения обрабатываются группами, что позволяет учитывать информацию о соседних коэффициентах. В модели с аддитивным шумом проведен анализ несмещенной оценки среднеквадратичного риска и показано, что при определенных условиях регулярности эта оценка является сильно состоятельной и асимптотически нормальной. Данные свойства позволяют использовать оценку риска в качестве критерия качества метода и строить асимптотические доверительные интервалы для теоретического среднеквадратичного риска.

Ключевые слова: вейвлеты; блочная пороговая обработка; несмещенная оценка риска; асимптотическая нормальность; сильная состоятельность

DOI: 10.14357/19922264210205

1 Введение

Методы вейвлет-анализа успешно применяются в задачах непараметрического оценивания функций и задачах обработки и анализа сигналов и изображений. Они приобрели свою популярность благодаря адаптивности, вычислительной эффективности и асимптотической оптимальности. Стандартные вейвлет-методы используют процедуры пороговой обработки, применяемые отдельно к каждому коэффициенту вейвлет-разложения. Коэффициент сравнивается с пороговым значением, и если его абсолютная величина оказывается меньше этого значения, то он обнуляется. Самыми популярными являются процедуры жесткой и мягкой пороговой обработки. Однако они имеют свои недостатки и часто не достигают оптимальных результатов. В частности, хорошо известный метод VisuShrink [1] приводит к слишком сглаженным оценкам функции. В работе [2] рассмотрен метод блочной пороговой обработки, при котором коэффициенты обрабатываются не отдельно, а группами. Цель такого подхода заключается в использовании информации о соседних коэффициентах. Получаемые оценки имеют оптимальный (в минимаксном смысле) порядок среднеквадратичного риска для различных классов функций [3].

Для практического анализа погрешности данного метода можно использовать несмещенную оценку среднеквадратичного риска [4], которая зависит только от наблюдаемых данных и дает возможность оценивать качество обработанного сигнала без использования «эталонных функций». В данной работе исследуются статистические свойства этой оценки и показывается, что для довольно широкого класса пространств Бесова она является асимптотически нормальной и сильно состоятельной. Для методов пороговой обработки отдельных коэффициентов подобные исследования проводились в работах [5–8] в предположении о принадлежности исследуемой функции сигнала классу равномерно регулярных по Липшицу функций.

2 Метод блочной пороговой обработки

Для функции сигнала $f \in L^2(\mathbb{R})$ разложение по вейвлет-базису имеет вид:

$$f = \sum_{j,k \in Z} \langle f, \psi_{j,k} \rangle \psi_{j,k}, \tag{1}$$

где $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$, а $\psi(t)$ — некоторая вейвлет-функция (семейство $\{\psi_{j,k}\}_{j,k \in Z}$ образует ортонормированный базис в $L^2(\mathbb{R})$). Индекс j

* Работа выполнена при финансовой поддержке РФФИ (проект 19-07-00352) и в соответствии с программой Московского центра фундаментальной и прикладной математики.

[1] Московский государственный университет имени М. В. Ломоносова, кафедра математической статистики факультета вычислительной математики и кибернетики; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, oshestakov@cs.msu.ru

в (1) называется масштабом, а индекс k — сдвигом. Преимущество разложения (1) заключается в «экономном» представлении функций, т. е. для довольно широкого класса функций лишь относительно небольшое число коэффициентов в (1) заметно отлично от нуля.

Если вейвлет-функция ψ имеет r непрерывных производных и r нулевых моментов, определим при $0 < \gamma < r$ и $1 \leqslant p, q \leqslant \infty$ полунорму последовательности вейвлет-коэффициентов выражением

$$|f|_{B^{\gamma}_{p,q}} = \left(\sum_{j=0}^{\infty} \left(2^{sj} \left(\sum_{k} |\langle f, \psi_{j,k} \rangle|^p \right)^{1/p} \right)^q \right)^{1/q},$$

где

$$s = \gamma + \frac{1}{2} - \frac{1}{p}.$$

Далее будем считать, что f задана на конечном отрезке и принадлежит пространству Бесова $B^{\gamma}_{p,q}(A)$ $(A > 0)$, т. е. $|f|_{B^{\gamma}_{p,q}} \leqslant A$.

Модель данных, рассматриваемая в данной работе, предполагает, что функция сигнала задана в дискретных отсчетах и наблюдения содержат шум:

$$X_i = f_i + \epsilon_i, \quad i = 1, \ldots, 2^J,$$

где 2^J — число отсчетов функции сигнала; f_i — незашумленные значения функции сигнала; ϵ_i — независимые нормально распределенные случайные величины с нулевым средним и дисперсией σ^2. После применения дискретного вейвлет-преобразования получается следующая модель зашумленных вейвлет-коэффициентов:

$$Y_{j,k} = \mu_{j,k} + \epsilon^W_{j,k}, \;\; j = 0, \ldots, J-1, \; k = 0, \ldots, 2^j - 1,$$

где $\epsilon^W_{j,k}$ независимы и имеют такое же распределение, как и ϵ_i, а $\mu_{j,k} \approx 2^{J/2} \langle f, \psi_{j,k} \rangle$.

Для подавления шума коэффициенты $Y_{j,k}$ подвергаются некоторой обработке. Самыми распространенными стали методы жесткой и мягкой пороговой обработки и их модификации [1, 9–16]. При использовании этих методов происходит сравнение абсолютной величины каждого коэффициента с некоторым порогом, и если это значение оказывается меньше порога, то коэффициент считается шумом и обнуляется. Такие методы обрабатывают каждый коэффициент отдельно, не используя информацию о других коэффициентах. Блочная пороговая обработка применяется к группам соседних коэффициентов, т. е. решение об обнулении принимается сразу по всем коэффициентам из группы.

Пусть $B_{j,1}, \ldots, B_{j,M_j}$ — разбиение множества индексов $\{0, \ldots, 2^j - 1\}$ на блоки одинаковой длины L (для удобства предположим, что 2^j делится на L). Пусть $S^2_{j,m} = \sum_{k \in B_{j,m}} Y^2_{j,k}$. Оценки коэффициентов $\mu_{j,k}$ вычисляются по правилу

$$\widehat{\mu}_{j,k} = \left(1 - \frac{TL\sigma^2}{S^2_{j,m}} \right)_+ Y_{j,k}, \; j = 0, \ldots, J-1, \; k \in B_m,$$

т. е. если величина $\sum_{k \in B_{j,m}} Y^2_{j,k}$ меньше порога $TL\sigma^2$, то все коэффициенты в рассматриваемом блоке обнуляются.

На качество оценок, получаемых с помощью блочной пороговой обработки, естественно влияют размер блока L и значение порога T. В работе [3] показано, что при $L = \log 2^J$ достигается баланс между локальной и глобальной адаптивностью метода блочной пороговой обработки, и если при этом $T^* \approx 4{,}50524$ (T^* является корнем уравнения $T - \log T - 3 = 0$), то среднеквадратичный риск оказывается (почти) оптимальным (в минимаксном смысле). В данной работе рассматриваются именно такие значения L и T.

3 Несмещенная оценка среднеквадратичного риска и ее свойства

Среднеквадратичный риск описанного выше метода определяется по формуле:

$$R_J(T) = \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \mathsf{E} \left(\widehat{\mu}_{j,k} - \mu_{j,k} \right)^2.$$

Вычислить его значение на практике нельзя, поскольку $R_J(T)$ зависит от ненаблюдаемых «чистых» коэффициентов $\mu_{j,k}$.

В [3] показано, что

$$\sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \mathsf{E} \left(\widehat{\mu}_{j,k} - \mu_{j,k} \right)^2 = \sum_{j=0}^{J-1} \sum_{m=1}^{M_j} \mathsf{E} F_m[T, L],$$

где

$$F_m(T, L) = \big[L\sigma^2 + \\ + \frac{T^2 L^2 \sigma^4 - 2TL\sigma^4 (L-2)}{S^2_{j,m}} \mathbb{1}(S^2_{j,m} > TL\sigma^2) + \\ + \left(S^2_{j,m} - 2L\sigma^2 \right) \mathbb{1} \left(S^2_{j,m} \leqslant TL\sigma^2 \right) \big].$$

Таким образом, величина

$$\widehat{R}_J(T) = \sum_{j=0}^{J-1} \sum_{m=1}^{M_j} F_m(T, L) \qquad (2)$$

служит несмещенной оценкой $R_J(T)$, не зависящей от $\mu_{j,k}$.

Покажем, что оценка (2) является асимптотически нормальной.

Теорема 1. *Пусть $f \in \underline{B}^{\gamma}_{p,q}(A)$ и задана на конечном отрезке. Пусть вейвлет-функция ψ имеет r непрерывных производных и r нулевых моментов, $r > \gamma$. Если $1 \leqslant p, q \leqslant \infty$ и $\gamma > \max(0, 1/p - 1/2)$, то при $J \to \infty$*

$$\mathsf{P}\left(\frac{\widehat{R}_J(T^*) - R_J(T^*)}{\sigma^2\sqrt{2^{J+1}}} < x\right) \to \Phi(x), \qquad (3)$$

где $\Phi(x)$ — функция распределения стандартного нормального закона.

Доказательство. Выберем $0 < d < 1$ и запишем:

$$\widehat{R}_J(T^*) - R_J(T^*) = \\ = \sum_{j=0}^{[dJ]}\sum_{m=1}^{M_j}\left[F_m(T^*, L) - \mathsf{E}F_m(T^*, L)\right] + \\ + \sum_{j=[dJ]+1}^{J-1}\sum_{m=1}^{M_j}\left[F_m(T^*, L) - \mathsf{E}F_m(T^*, L)\right]. \qquad (4)$$

Число слагаемых в первой сумме не превосходит $2^{[dJ]+1}$. Кроме того, существует такая константа $C_F > 0$, что

$$|F_m(T^*, L) - \mathsf{E}F_m(T^*, L)| \leqslant C_F T^* L \ \text{ п. в.} \qquad (5)$$

Применяя неравенство Хеффдинга, получаем, что для любого $\delta > 0$ найдется константа $C_\delta > 0$ такая, что

$$\mathsf{P}\left(\left|\sum_{j=0}^{[dJ]}\sum_{m=1}^{M_j}\left[F_m(T^*, L) - \right.\right.\right. \\ \left.\left.\left. - \mathsf{E}F_m(T^*, L)\right]\right| \Big/ \left(\sigma^2\sqrt{2^{J+1}}\right) > \delta\right) \leqslant \\ \leqslant \exp\left\{-C_\delta 2^{J-dJ}\right\}, \qquad (6)$$

т. е.

$$\frac{\sum_{j=0}^{[dJ]}\sum_{m=1}^{M_j}\left[F_m(T^*, L) - \mathsf{E}F_m(T^*, L)\right]}{\sigma^2\sqrt{2^{J+1}}} \xrightarrow{\mathsf{P}} 0 \\ \text{при } J \to \infty. \qquad (7)$$

При $p_1 \leqslant p_2$ справедливы неравенства [3]:

$$\left(\sum_{k=0}^{2^j-1}|\mu_{j,k}|^{p_2}\right)^{1/p_2} \leqslant \left(\sum_{k=0}^{2^j-1}|\mu_{j,k}|^{p_1}\right)^{1/p_1} \leqslant \\ \leqslant 2^{j(1/p_1 - 1/p_2)}\left(\sum_{k=0}^{2^j-1}|\mu_{j,k}|^{p_2}\right)^{1/p_2}. \qquad (8)$$

Так как $f \in B^{\gamma}_{p,q}(A)$, то

$$2^{js}\left(\sum_{k=0}^{2^j-1}|\mu_{j,k}|^{p}\right)^{1/p} \leqslant A2^{J/2}$$

и из неравенств (8) следует, что при $p \geqslant 2$

$$\sum_{k=0}^{2^j-1}\mu_{j,k}^2 \leqslant A^2 2^{-2\gamma j + J}. \qquad (9)$$

Для произвольного $\varepsilon > 0$ при всех $j > dJ$ существует не более $A^2 2^{J - 2\gamma j + \varepsilon j}$ слагаемых в (9) таких, что $\sum_{k\in B_{j,m}}\mu_{j,k}^2 > 2^{-\varepsilon j}$. Выделяя эти слагаемые из второй суммы в (4) в отдельную сумму S' и применяя к S' неравенство Хеффдинга, аналогичное (6), получаем, что

$$\frac{S'}{\sigma^2\sqrt{2^{J+1}}} \xrightarrow{\mathsf{P}} 0 \text{ при } J \to \infty.$$

Таким образом, без ограничения общности можно считать, что во второй сумме в (4) $\sum_{k\in B_{j,m}}\mu_{j,k}^2 \to \\ \to 0$ при $J \to \infty$ для всех $m = 1, \ldots, M_j$ и $j > dJ$.

Если $p < 2$ и $\gamma > 1/p - 1/2$, то $s > 0$ и из неравенств (8) следует, что

$$\sum_{k=0}^{2^j-1}\mu_{j,k}^2 \leqslant A^2 2^{-2sj + J}. \qquad (10)$$

Рассуждая аналогично случаю $p \geqslant 2$, заключаем, что без ограничения общности можно считать, что во второй сумме в (4) также $\sum_{k\in B_{j,m}}\mu_{j,k}^2 \to 0$ при $J \to \infty$ для всех $m = 1, \ldots, M_j$ и $j > dJ$.

Рассмотрим дисперсии слагаемых во второй сумме в (4). Имеем

$$\mathsf{D}F_m(T^*, L) = \mathsf{D}\left[S_{j,m}^2 - L\sigma^2 + U_{j,m}(T^*, L)\right],$$

где

$$U_{j,m}(T^*, L) = \\ = \left[\frac{(T^*L)^2\sigma^4 - 2T^*L\sigma^4(L-2)}{S_{j,m}^2} - S_{j,m}^2 + 2L\sigma^2\right] \times \\ \times \mathbb{1}(S_{j,m}^2 > T^*L\sigma^2).$$

Пусть X_L — случайная величина, имеющая распределение χ^2_L с L степенями свободы, и C_1 и C_2 —

некоторые положительные константы. Тогда в силу (9) или (10) для некоторой константы $1 < C_\chi < 1 + \delta_\chi$ $(0 < \delta_\chi < 1)$

$$\begin{aligned}\mathsf{D}\left[S_{j,m}^2 \mathbb{1}\left(S_{j,m}^2 > T^* L\sigma^2\right)\right] \leqslant \\ \leqslant \mathsf{E}\left[\left(S_{j,m}^2\right)^2 \mathbb{1}(S_{j,m}^2 > T^* L\sigma^2)\right] \leqslant \\ \leqslant C_\chi^2 \mathsf{E}\left[X_L^2 \mathbb{1}(C_\chi X_L > T^* L)\right] \leqslant \\ \leqslant C_1 (T^* L)^2 \mathsf{P}\left(C_\chi X_L > T^* L\right)\end{aligned}$$

и аналогично

$$\mathsf{D}\left(\left[\frac{(T^* L)^2\sigma^4 - 2T^* L\sigma^4(L-2)}{S_{j,m}^2} + 2L\sigma^2\right] \times \right.$$

$$\left. \times \mathbb{1}(S_{j,m}^2 > T^* L\sigma^2)\right) \leqslant C_2 (T^* L)^2 \mathsf{P}\left(C_\chi X_L > T^* L\right).$$

В силу леммы 2 из работы [3] и вида T^* и L

$$(T^* L)^2 \mathsf{P}\left(C_\chi X_L > T^* L\right) \to 0 \text{ при } J \to \infty\,.$$

Таким образом, учитывая очевидное соотношение для дисперсии разности случайных величин $X - Y$:

$$\left(\sqrt{\mathsf{D}X} - \sqrt{\mathsf{D}Y}\right)^2 \leqslant \mathsf{D}(X - Y) \leqslant \left(\sqrt{\mathsf{D}X} + \sqrt{\mathsf{D}Y}\right)^2,$$

получаем

$$\lim_{J\to\infty} \frac{\mathsf{D}\sum_{j=[dJ]+1}^{J-1} \sum_{m=1}^{M_j} F_m(T^*, L)}{\mathsf{D}\sum_{j=[dJ]+1}^{J-1} \sum_{k=0}^{2^j-1} Y_{j,k}^2} = 1\,. \quad (11)$$

Кроме того, поскольку $Y_{j,k}$ независимы и $\mathsf{D}Y_{j,k}^2 = 2\sigma^4 + 4\sigma^2\mu_{j,k}^2$, получаем

$$\lim_{J\to\infty} \frac{\mathsf{D}\sum_{j=[dJ]+1}^{J-1} \sum_{k=0}^{2^j-1} Y_{j,k}^2}{\sigma^4 2^{J+1}} = 1\,. \quad (12)$$

Наконец, выполнено условие Линдеберга: для любого $\epsilon > 0$ при $J \to \infty$

$$\frac{1}{D_J^2} \sum_{j=[dJ]+1}^{J-1} \sum_{m=1}^{M_j} \mathsf{E}\big[(F_m(T^*, L) - \mathsf{E}F_m(T^*, L))^2 \times$$

$$\times \mathbb{1}\left(|F_m(T^*, L) - \mathsf{E}F_m(T^*, L)| > \epsilon D_J\right)\big] \to 0\,, \quad (13)$$

где

$$D_J^2 = \mathsf{D} \sum_{j=[dJ]+1}^{J-1} \sum_{m=1}^{M_j} \left[F_m(T^*, L) - \mathsf{E}F_m(T^*, L)\right].$$

Действительно, в силу (5), (11) и (12) начиная с некоторого J все индикаторы в (13) обращаются в ноль. Объединяя (7), (12) и (13), получаем (3). Теорема доказана.

Докажем теперь свойство сильной состоятельности оценки (2), справедливое при более слабых ограничениях.

Теорема 2. *Пусть $f \in L^2(\mathbb{R})$ и задана на конечном отрезке, тогда при любом $\alpha > 1/2$ имеет место сходимость*

$$\frac{\widehat{R}_J(T^*) - R_J(T^*)}{2^{\alpha J}} \to 0 \text{ п. в. при } J \to \infty\,. \quad (14)$$

Доказательство. Используя неравенство Хеффдинга, с учетом (5) получаем, что для любого $\delta > 0$ найдется константа $C_\delta > 0$ такая, что

$$p_J = \mathsf{P}\left(\frac{\left|\widehat{R}_J(T^*) - R_J(T^*)\right|}{2^{\alpha J}} > \delta\right) \leqslant$$

$$\leqslant \exp\left\{-C_\delta 2^{2\alpha J - J}\right\},$$

и, поскольку $\sum_{J=1}^{\infty} p_J < \infty$, в силу леммы Бореля–Кантелли выполнено (14). Теорема доказана.

Теоремы 1 и 2 дают теоретическое обоснование использования значения $\widehat{R}_J(T^*)$ в качестве оценки неизвестной величины риска (погрешности) $R_J(T^*)$, а также дают возможность строить асимптотические доверительные интервалы для $R_J(T^*)$.

Литература

1. *Donoho D., Johnstone I. M.* Ideal spatial adaptation via wavelet shrinkage // Biometrika, 1994. Vol. 81. No. 3. P. 425–455.
2. *Hall P., Kerkyacharian G., Picard D.* On the minimax optimality of block thresholded wavelet estimators // Stat. Sinica, 1999. Vol. 9. P. 33–50.
3. *Cai T.* Adaptive wavelet estimation: A block thresholding and oracle inequality approach // Ann. Stat., 1999. Vol. 28. No. 3. P. 898–924.
4. *Stein C.* Estimation of the mean of a multivariate normal distribution // Ann. Stat., 1981. Vol. 9. No. 6. P. 1135–1151.
5. *Шестаков О. В.* Асимптотическая нормальность оценки риска пороговой обработки вейвлет-коэффициентов при выборе адаптивного порога // Докл. Акад. наук, 2012. Т. 445. № 5. С. 513–515.
6. *Shestakov O. V.* On the strong consistency of the adaptive risk estimator for wavelet thresholding // J. Math. Sci., 2016. Vol. 214. No. 1. P. 115–118.
7. *Шестаков О. В.* Статистические свойства метода подавления шума, основанного на стабилизированной жесткой пороговой обработке // Информатика и её применения, 2016. Т. 10. Вып. 2. С. 65–69.
8. *Попенова П. С., Шестаков О. В.* Анализ статистических свойств метода гибридной пороговой обработки // Вестн. Тверск. ун-та. Сер.: Прикладная математика. 2019. № 1. С. 15–22.

9. *Donoho D., Johnstone I. M.* Adapting to unknown smoothness via wavelet shrinkage // J. Am. Stat. Assoc., 1995. Vol. 90. P. 1200–1224.
10. *Donoho D., Johnstone I. M.* Minimax estimation via wavelet shrinkage // Ann. Stat., 1998. Vol. 26. No. 3. P. 879–921.
11. *Gao H.-Y.* Wavelet shrinkage denoising using the non-negative garrote // J. Comput. Graph. Stat., 1998. Vol. 7. No. 4. P. 469–488.
12. *Poornachandra S., Kumaravel N., Saravanan T. K., Somaskandan R.* WaveShrink using modified hyper-shrinkage function // 27th Annual Conference (International) of the IEEE Engineering in Medicine and Biology Society Proceedings. — Piscataway, NJ, USA: IEEE, 2005. P. 30–32.
13. *Lin Y., Cai J.* A new threshold function for signal denoising based on wavelet transform // Conference (International) on Measuring Technology and Mechatronics Automation Proceedings. — Piscataway, NJ, USA: IEEE, 2010. P. 200–203.
14. *Huang H.-C., Lee T. C. M.* Stabilized thresholding with generalized sure for image denoising // 17th Conference (International) on Image Processing Proceedings. — Piscataway, NJ, USA: IEEE, 2010. P. 1881–1884.
15. *He C., Xing J., Li J., Yang Q., Wang R.* A new wavelet thresholding function based on hyperbolic tangent function // Math. Probl. Eng., 2015. Vol. 2015. Art. 528656.
16. *Zhao R.-M., Cui H.-M.* Improved threshold denoising method based on wavelet transform // 7th Conference (International) on Modelling, Identification and Control Proceedings. — Piscataway, NJ, USA: IEEE, 2015. Art. 7409352. 4 p. doi: 10.1109/ICMIC.2015.7409352.
17. *Mallat S.* A wavelet tour of signal processing. — New York, NY, USA: Academic Press, 1999. 857 p.

Поступила в редакцию 27.03.2021

ANALYSIS OF THE UNBIASED MEAN-SQUARE RISK ESTIMATE OF THE BLOCK THRESHOLDING METHOD

O. V. Shestakov[1,2]

[1]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[2]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Signal and image processing methods based on wavelet decomposition and thresholding have become very popular in solving problems of compression and noise suppression. This is due to their ability to adapt to local features of functions, high speed of processing algorithms and optimality of estimates obtained. In this paper, a block thresholding method is considered, in which expansion coefficients are processed in groups, which makes it possible to take into account information about neighboring coefficients. In the model with additive noise, an unbiased estimate of the mean-square risk is analyzed and it is shown that, under certain conditions of regularity, this estimate is strongly consistent and asymptotically normal. These properties allow using the risk estimate as a quality criterion for the method and constructing asymptotic confidence intervals for the theoretical mean-square risk.

Keywords: wavelets; block thresholding; mean-square risk estimate; asymptotic normality; strong consistency

DOI: 10.14357/19922264210205

Acknowledgments

This research was supported by the Russian Foundation for Basic Research (project 19-07-00352). The research was conducted in accordance with the program of the Moscow Center for Fundamental and Applied Mathematics.

References

1. Donoho, D., and I. M. Johnstone. 1994. Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81(3):425–455.
2. Hall, P., G. Kerkyacharian, and D. Picard. 1999. On the minimax optimality of block thresholded wavelet estimators. *Stat. Sinica* 9:33–50.
3. Cai, T. 1999. Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Stat.* 28(3):898–924.
4. Stein, C. 1981. Estimation of the mean of a multivariate normal distribution. *Ann. Stat.* 9(6):1135–1151.
5. Shestakov, O. V. 2012. Asymptotic normality of adaptive wavelet thresholding risk estimation. *Dokl. Math.* 86(1):556–558.

6. Shestakov, O. V. 2016. On the strong consistency of the adaptive risk estimator for wavelet thresholding. *J. Math. Sci.* 214(1):115–118.
7. Shestakov, O. V. 2016. Statisticheskie svoystva metoda podavleniya shuma, osnovannogo na stabilizirovannoy zhestkoy porogovoy obrabotke [Statistical properties of the denoising method based on the stabilized hard thresholding]. *Informatika i ee Primeneniya — Inform. Appl.* 10(2):65–69.
8. Popenova, P. S., and O. V. Shestakov. 2019. Analiz statisticheskikh svoystv metoda gibridnoy porogovoy obrabotki [Analysis of statistical properties of the hybrid thresholding technique]. *Vestnik Tverskogo gosudarstvennogo un-ta. Ser. Prikladnaya matematika* [Bull. of the Tverskoy State University. Ser. Appl. Math.] 1:15–22.
9. Donoho, D., and I. M. Johnstone. 1995. Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* 90:1200–1224.
10. Donoho, D., and I. M. Johnstone. 1998. Minimax estimation via wavelet shrinkage. *Ann. Stat.* 26(3):879–921.
11. Gao, H.-Y. 1998. Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Stat.* 7(4):469–488.
12. Poornachandra, S., N. Kumaravel, T. K. Saravanan, and R. Somaskandan. 2005. WaveShrink using modified hyper-shrinkage function. *27th Annual Conference (International) of the IEEE Engineering in Medicine and Biology Society Proceedings*. Piscataway, NJ: IEEE. 30–32.
13. Lin, Y., and J. Cai. 2010. A new threshold function for signal denoising based on wavelet transform. *Conference (International) on Measuring Technology and Mechatronics Automation Proceedings*. Piscataway, NJ: IEEE. 200–203.
14. Huang, H.-C., and T. C. M. Lee. 2010. Stabilized thresholding with generalized sure for image denoising. *17th Conference (International) on Image Processing Proceedings*. Piscataway, NJ: IEEE. 1881–1884.
15. He, C., J. Xing, J. Li, Q. Yang, and R. Wang. 2015. A new wavelet thresholding function based on hyperbolic tangent function. *Math. Probl. Eng.* 2015:528656. 10 p.
16. Zhao, R.-M., and H.-M. Cui. 2015. Improved threshold denoising method based on wavelet transform. *7th Conference (International) on Modelling, Identification and Control Proceedings*. Piscataway, NJ: IEEE. Art. ID: 7409352. 4 p. doi: 10.1109/ICMIC.2015.7409352.
17. Mallat, S. 1999. *A wavelet tour of signal processing*. New York, NY: Academic Press. 857 p.

Received March 27, 2021

Contributor

Shestakov Oleg V. (b. 1976) — Doctor of Science in physics and mathematics, professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; oshestakov@cs.msu.su

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ПОПОЛНЯЕМЫХ КОЛЛЕКЦИЙ BIG DATA В РЕЖИМЕ ПРОЦЕССНО-РЕАЛЬНОГО ВРЕМЕНИ*

А. А. Грушо[1], М. И. Забежайло[2], Д. В. Смирнов[3], Е. Е. Тимонина[4]

Аннотация: Обсуждается задача поиска релевантных заданной цели данных в постоянно пополняемых новой информацией коллекциях Big Data в условиях жестких ограничений на допустимое время (так называемое процессно-реальное время) анализа данных (АД) и поддержки принятия решений (ППР). В основе развиваемого подхода — использование современных методов искусственного интеллекта, в частности представления знаний и формализации рассуждений в системах интеллектуального АД (ИАД). Рассматривается ряд критически значимых для результативности такого ИАД барьеров, в том числе обусловленных доказуемой трудноразрешимостью возникающих здесь комбинаторных задач, особенностями представления знаний и управления перебором вариантов, а также некоторыми аспектами управления полнотой и точностью порождаемых результатов. Представлена схема формализации развиваемой процедурной конструкции ИАД. Обсуждаемый подход сопровождается иллюстрациями его реализации в рамках системы идентификации признаков вредоносной инсайдерской активности в крупном отечественном коммерческом банке.

Ключевые слова: Big Data; процессно-реальное время; интеллектуальный анализ данных; информационная безопасность; поиск инсайдеров

DOI: 10.14357/19922264210206

1 Введение

Потребность в разработке проблемно-ориентированных средств «навигации» в Big Data очевидным образом ассоциируется с потребностью в разработке эффективных поисковых технологий, которые позволяли бы результативно обрабатывать большие объемы не только собственно исходных «сырых» данных, но и формируемой на их основе аналитики — витрин данных, графической информации, dashboard'ов и др. Критичными здесь оказываются потребности в обработке больших объемов постоянно изменяющейся, пополняемой новыми сведениями информации в условиях жестких временны́х ограничений (так называемого процессно-реального времени) АД и ППР.

При решении поисковых задач такого типа сегодня широко используется ряд коммерческих корпоративных поисковых систем, в частности Algolia [1], IBM Watson Discovery [2], Yext [3], Swiftype [4], SearchUnify [5]. Популярны корпоративные поисковые системы с открытым исходным кодом, например ElasticSearch [6], Solr [7], Sphinx [8]. Среди отечественных коммерческих корпоративных поисковых систем, по-видимому, наиболее известны Спутник (Ростелеком) [9] и 1С [10].

Говоря о критически значимых характеристиках корпоративных поисковых систем (так называемых *поисковиков* — см. [1–10]), следует в первую очередь обратить внимание:

- на скорость индексирования первичных данных, т. е. быстроту переработки поисковиком входных «сырых» данных для занесения в свой внутренний поисковый аппарат — системы поисковых индексов, классификаторы и т. п. Обычно этот параметр оценивается в мегабайтах чистого «сырого» входного текста в секунду;
- на скорость переиндексации — реконструкции поискового инструментария, т. е. обновления индексов или создания новых с приходом новой входной информации. При этом может поддерживаться как инкрементальное индексирование, так и полная перестройка (переформирование) индекса, могут использоваться и дополнительные индексы, в том числе так называемый дельта-индекс, в который включается только новая информация;

*Работа частично поддержана РФФИ (проект 18-29-03081).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, grusho@yandex.ru

[2]Вычислительный центр Федерального исследовательского центра «Информатика и управление» Российской академии наук, m.zabezhailo@yandex.ru

[3]ПАО Сбербанк России, dvlsmirnov@sberbank.ru

[4]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, eltimon@yandex.ru

– на поддерживаемые API (application programming interface). Поисковое ядро необходимо связывать с приложениями, которые могут иметь библиотеки, работающие с API поисковика;

– на взаимосвязь размеров базы и скорости поиска, так как некоторые поисковики попросту перестают отвечать на запросы при индексах, содержащих более 50 млн записей;

– на поддерживаемые типы входных документов, т. е. возможность индексации различных типов источников — систем управления базами данных, файловых хранилищ и т. п.

К сожалению, на текущий момент рынок не предлагает надежных коммерческих систем поддержки поиска и идентификации признаков вредоносной инсайдерской активности, способных обеспечить результативное применение в крупных отечественных финансовых структурах. Именно по этим причинам целенаправленного обсуждения заслуживает разработка методики выявления признаков инсайдерской активности и создание компьютерной системы поддержки профильной деятельности оперативных работников служб безопасности.

2 Профиль угроз

Стандартный подход к решению задачи идентификации признаков вредоносной инсайдерской активности в текущих данных мониторинга функционирования объекта защиты базируется на выделении в наблюдаемых данных отслеживаемого поведения пользователей защищаемой системы таких фактических действий, которые могут быть классифицированы как потенциально опасное, способное привести к вредоносным последствиям поведение.

Разработка методики и инструментария идентификации признаков инсайдерской активности основана на формировании актуальной модели угроз. Модель угроз формализуется в виде профиля угроз (ПУ), представляющего собой постоянно поддерживаемый в актуальном состоянии перечень так называемых типовых сценариев (ТС). Исходные ТС рождаются из информации, взятой из опыта оперативных сотрудников, вовлеченных в расследования конкретных случаев мошенничества (признаки инсайдерской активности). Опыт оперативных сотрудников сначала фиксируется в виде текстового описания, которое далее преобразуется в машиночитаемый формализованный вид. При этом может быть задействовано промежуточное представление знаний о каждом из ТС в виде фрейма. Для описания данных в слотах подобных фреймов предусмотрены иерархии типов данных: от булевых значений признаков: «да»/«нет» — до графов параметров и отношений между такими параметрами с пометками на вершинах и ребрах, а также текстовых комментариев, например в виде Binary Large Objects (BLOB). Кроме обобщения опыта в формировании ТС используется ИАД и машинное обучение (МО).

Простейший вариант представления знаний в ТС ПУ — использование булевых значений «да»/«нет», позволяющих описать каждый такой фрейм в виде множества характеризующих именно его признаков. В свою очередь, множество всех используемых при описании текущего ПУ признаков определяет битовый вектор, соответствующими единицами которого кодируется каждый из соответствующих ТС в ПУ. Обработка машиночитаемого описания фреймов, представленных в виде битовых векторов, дает возможность получить существенный выигрыш в производительности при анализе текущих данных, так как позволяет организовать сравнение текущей ситуации с описаниями ТС средствами одной вычислительной макрооперации.

Множество ТС представляет систему фильтров, позволяющих сократить объем данных для дальнейшего анализа и, по возможности, не пропустить информацию о признаках действий инсайдеров.

3 Организация быстрой фильтрации на основе множества типовых сценариев

Критически важную роль в обеспечении результативности обсуждаемого подхода к выделению из исходных «сырых» Big Data информации, содержащей признаки вредоносной инсайдерской активности, играют две группы характеристик:

(1) *чувствительность* критериев релевантности ПУ, т. е. возможности ошибок первого и второго рода и, как следствие, возможности неполноты идентификации признаков инсайдерской активности и ложных срабатываний;

(2) *размерность* текущего описания ПУ, определяющая параметры перебора всех совпадений его фрагментов с элементами описания текущего множества отслеживаемых данных.

Если перебирать совпадения описаний каждого ТС с текущим множеством отслеживаемых данных по каждому пользователю, то задача приобретет

экспоненциальную сложность. Однако даже в этой ситуации перебор можно сократить за счет отказа от повторных проверок общих для нескольких ТС фрагментов:

- выделения всех общих для всех ТС частей;
- упорядочения их по взаимной вложимости, т. е. формирования частично упорядоченного множества таких фрагментов в виде диаграммы сходств;
- организации выделения общих с множеством отслеживаемых данных фрагментов, начиная с нижнего «этажа» этой диаграммы с наиболее часто встречаемых в ТС общих фрагментов, затем двигаясь «вверх» к наименее часто встречающимся общим фрагментам и далее — к собственно полным описаниям актуальных ТС.

В общем случае приходится иметь дело с экспоненциально быстро растущим (с линейным ростом размеров описания текущих ТС) числом элементов в такой диаграмме. Однако можно показать, что верхняя и нижняя границы такой диаграммы могут быть построены полиномиально быстро, а анализ большинства реальных множеств отслеживаемых данных требует проверки лишь нескольких из всех цепей частичного порядка в таких диаграммах. Каждая из таких цепей, ведущая от одного из элементов нижней границы диаграммы к одному из ее верхних элементов, имеет длину, ограниченную полиномом от размеров текущего описания имеющихся ТС. Это легко объясняется содержательными соображениями, т. е. характеристиками бизнес-активности пользователей в реальных ситуациях. Дополнительно «навигация» «снизу вверх» по цепям частичного порядка открывает возможности проактивной ориентации офицеров безопасности в «подсвеченных» ситуациях, демонстрирующих релевантность текущей отслеживаемой ситуации фрагментам описания какого-либо из известных ТС. Таким образом, этот метод позволяет целенаправленно фокусировать внимание и ресурсы на потенциально опасных направлениях развития каждой конкретной отслеживаемой ситуации в динамике ее изменений.

4 Формирование типовых сценариев с помощью интеллектуального анализа данных и машинного обучения

Первоначально ТС строятся на основе опыта оперативных работников, т. е. ТС формируются как прецеденты, содержащие признаки инсайдеров (они помечаются меткой «+»), и как прецеденты, которые заведомо не связаны с деятельностью инсайдеров (помечаются меткой «−»). Остальные прецеденты помечаются меткой «0» как неопределенные. Описания прецедентов формализуются следующим образом.

Исходные («сырые») данные описываются множеством наблюдаемых значений параметров $x_1, x_2, \ldots, x_n$. В условиях открытости это множество может меняться. Каждый параметр x_i, $i = 1, \ldots, n$, имеет домен своих значений (характеристик):

$$A_1 = \{a_{11}, a_{12}, \ldots, a_{1m_1}\};$$
$$A_2 = \{a_{21}, a_{22}, \ldots, a_{2m_2}\};$$
$$\ldots$$
$$A_n = \{a_{n1}, a_{n2}, \ldots, a_{nm_n}\}.$$

Одним из элементов каждого из этих множеств является «пустой» символ, который означает, что соответствующий параметр не участвует в изучаемом объекте.

Определение 1. Объектом o называется произвольный элемент множества $A_1 \times A_2 \times \cdots \times A_n$.

Формализация опыта оперативных работников выражается в формировании классов:

$$O^+ = \{o_1^+, o_2^+, \ldots, o_s^+\} \subseteq A_1 \times A_2 \times \cdots \times A_n = \prod_{i=1}^{n} A_i;$$

$$O^- = \{o_1^-, o_2^-, \ldots, o_r^-\} \subseteq \prod_{i=1}^{n} A_i;$$

$$O^0 = \{o_1^0, o_2^0, \ldots, o_v^0\} \subseteq \prod_{i=1}^{n} A_i.$$

Определим множество ТС на обученных прецедентах как множество векторов $O^+ \cup O^-$.

Если описывать ТС с помощью длинных двоичных векторов, то можно использовать возможности вычислительной техники для быстрого и качественного отбора прецедентов в соответствии с построенной классификацией. Множество $O^+ \cup O^-$ объявляется релевантным на данном этапе и обозначается R.

Определим функцию f_R: $A_1 \times A_2 \times \cdots \times A_n \to \to \{0, 1\}$, где

$$f_R = \begin{cases} 1, & o \in R; \\ 0, & o \notin R. \end{cases}$$

Обозначим через $A \subseteq \prod_{i=1}^{n} A_i$ текущее множество «сырых» данных.

Задача расширения множества ТС состоит в доопределении функции f_R (где это возможно) на

данные $A\backslash(O^+ \cup O^-)$. Доопределение значений функции f_R на исходно неопределенных элементах множества $A\backslash(O^+ \cup O^-)$ может быть выполнено по традиционной для МО (*интерполяционно-экстраполяционной*) схеме.

I этап. На элементах множества $O^+ \cup O^-$ строятся *зависимости* того или иного класса таким образом, что все кортежи из множества O^+ «лежат» на каких-либо из таких зависимостей, а все кортежи из множества O^- не «лежат» ни на одной из них. При этом на каждом элементе множества O^+ хотя бы одна из таких зависимостей выполняется, а на каждом из элементов множества O^- не выполняется ни одна из таких зависимостей.

II этап. Построенные на первом этапе зависимости экстраполируются, где это возможно, на исходно недоопределенные кортежи из множества $A\backslash(O^+ \cup O^-)$. Все такие случаи классифицируются как $f_R = 1$, а во всех оставшихся случаях $f_R = 0$.

Процесс экстраполяции отношения R с имеющихся примеров и контрпримеров на «сырые» данные — процедура восстановления значений функции f_R на оставшихся кортежах из $A\backslash(O^+ \cup \cup O^-)$. Она основана на анализе сходства описаний прецедентов, уточняемого как бинарная алгебраическая операция [11]. При этом

(а) задействовано описание прецедентов в виде кортежа значений соответствующих параметров $x_1, x_2, \ldots, x_n$;

(б) на $x_1, x_2, \ldots, x_n$ определяется (покомпонентными сравнениями значений параметров) бинарная операция сходства $\otimes$, удовлетворяющая стандартным условиям [11]: $\forall\, o_1, o_2, o_3$ выполнено

(1) $o \otimes o = o$;

(2) $o_1 \otimes o_2 = o_2 \otimes o_1$;

(3) $o_1 \otimes o_2 \otimes o_3 = (o_1 \otimes o_2) \otimes o_3 = o_1 \otimes (o_2 \otimes o_3)$.

При этом значения параметра x_i у сравниваемых объектов o_1 и o_2 считаются сходными, если $x_i \in A_i^* \subseteq A_i$, где A_i^* — заранее заданная окрестность значений параметра x_i. Для признаков, принимающих лишь булевы значения 0 или 1, это будет операция пересечения множеств $\cap$;

(в) отношение сходства прецедентов определяется по непустому результату вычисления операции сходства представляющих эти прецеденты кортежей значений признаков. Два прецедента сходны, если результат применения операции $\otimes$ к их описаниям не является пустым объектом;

(г) для каждого прецедента o из множества A класс сходства $\mathrm{T}(o)$ всех сходных с ним прецедентов из A формируется объединением в $\mathrm{T}(o)$ всех элементов множества A, сходных с o;

(д) сформированные классы сходства после дополнительного анализа эмпирических закономерностей распадаются на две части, построенные на позитивных примерах и построенные на контрпримерах;

(е) для каждого из всех имеющихся позитивных, т. е. построенных на примерах, классов сходства $\mathrm{T}(o)$ выделяются подклассы $\mathrm{T}_V(o)$, каждый из которых порождается одним из соответствующих (участвующих в порождении этого $\mathrm{T}(o)$) сходств V примеров из $\mathrm{T}(o)$, т. е. V — множество значений вектора о, принадлежащих всем векторам из множества $\mathrm{T}_V(o)$. Из множества $\mathrm{T}_V(o)$ удаляются все такие элементы, в которые вкладывается хотя бы один из контрпримеров из множества O^-. Это условие проверяется по всем V, участвующим в порождении $\mathrm{T}(o)$;

(ж) экстраполяция частично-определенного отношения R на все элементы множества $A\backslash(O^+ \cup \cup O^-)$ реализуется с помощью проверки вложимости каждого элемента $A\backslash(O^+ \cup O^-)$ в какой-либо подкласс $\mathrm{T}_V(o)$ позитивных классов сходства $\mathrm{T}(o)$. В случае такого попадания $f_R = 1$ на данном прецеденте, в противном случае считается, что $f_R = 0$. Таким образом, расширяется множество ТС с признаками вредоносной активности инсайдера. Аналогично из множества элементов, для которых $f_R = 0$, можно выделять ТС с отрицательными признаками наличия вредоносной инсайдерской активности.

5 Примеры профилей угроз и диаграмма сходств типовых сценариев

Рассмотрим диаграммы D_{TC} сходств ТС на примере представления знаний в описывающих ТС фреймах в виде булевых векторов признаков. Множество V в построенных положительных классах $\mathrm{T}_V(o)$ назовем актуальными признаками ТС, принадлежащим $\mathrm{T}_V(o)$.

Тогда каждый ТС может быть представлен как кортеж из 0 и 1, в котором единицы соответствуют актуальным для данного ТС признакам.

Пример. Пусть $\mathrm{TC}_{\mathrm{AT}}$ — сценарий, текстовое описание которого фиксирует опасность одновременного доступа пользователя в аналитические (А) и транзакционные (Т) приложения. Тогда актуальные признаки сценария $\mathrm{TC}_{\mathrm{AT}}$:

P_1 — доступ в аналитический блок;

P_2 — доступ в транзакционный блок;

P_3 – идентификация имеющего эти доступы пользователя с учетом штатного профиля его доступов.

Таким образом, $\mathrm{TC}_{\mathrm{AT}}$ сопоставлено множество признаков $\{P_1, P_2, P_3\}$.

В построенном языке представления знаний актуальный ПУ представляет собой набор множеств актуальных для соответствующего ТС признаков, которые можно представить в виде набора векторов вида:

$$\alpha_{\mathrm{TC}_1} = \langle \alpha_{\mathrm{TC}_1,1}, \alpha_{\mathrm{TC}_1,2}, \dots \alpha_{\mathrm{TC}_1,n} \rangle ;$$
$$\alpha_{\mathrm{TC}_2} = \langle \alpha_{\mathrm{TC}_2,1}, \alpha_{\mathrm{TC}_2,2}, \dots \alpha_{\mathrm{TC}_2,n} \rangle ;$$
$$\dots$$
$$\alpha_{\mathrm{TC}_m} = \langle \alpha_{\mathrm{TC}_m,1}, \alpha_{\mathrm{TC}_m,2}, \dots \alpha_{\mathrm{TC}_m,n} \rangle ,$$

где $\text{ПУ} = \{\mathrm{TC}_1, \mathrm{TC}_2, \dots, \mathrm{TC}_m\}$.

Вычисление сходств между векторами α_{TC_i} реализуется с помощью бинарной операции $\otimes$ покомпонентного сравнения булевых векторов:

$$\alpha_{\mathrm{TC}_i,1} \otimes \alpha_{\mathrm{TC}_j,1} = 1 \Leftrightarrow \alpha_{\mathrm{TC}_i,1} = \alpha_{\mathrm{TC}_j,1} = 1 .$$

В остальных случаях результат сравнения равен 0.

Таким образом может быть сформирована диаграмма D_{TC} всех непустых сходств описаний ТС, имеющихся на текущий момент. При этом:

- нижний «этаж» этой диаграммы формируется сходствами (множествами общих признаков) максимальных по числу элементов возможных подмножеств ТС из всего текущего ПУ;
- верхний «этаж» этой диаграммы формируется минимальными по числу образующих их ТС и максимальными по числу общих признаков сходствами ТС.

Проверка «сырых» данных на возможное отношение к известным признакам инсайдерской активности, представленным описаниями ТС, начинается с нижнего «этажа» диаграммы D_{TC} и далее осуществляется не сравнением «всех» со «всеми», а лишь просмотром релевантных цепей частичного порядка (по взаимной вложимости сходств) «снизу вверх» (к «верхнему» этажу диаграммы D_{TC}, а далее — к полным описаниям соответствующих ТС).

6 Оценки сложности вычислений при формировании D_{TC}

Пусть заданы два множества: $A = \{a_1, a_2, \dots, a_n\}$ — множество признаков, используемых при формировании ТС из ПУ, и текущее множество $\text{ПУ} = \{\mathrm{TC}_1, \mathrm{TC}_2, \dots, \mathrm{TC}_m\} \subseteq 2^A \backslash \phi$ описаний ТС в виде подмножеств признаков из A. В [12, 13] показано, что справедливы следующие утверждения.

Утверждение 1. Задача о вычислении числа элементов диаграммы D_{TC} принадлежит классу # РС — так называемых перечислительно полных комбинаторных проблем [14–16].

Таким образом, в общем случае размер диаграммы D_{TC} растет *экспоненциально* быстро с линейным ростом размеров множеств A и ПУ.

Утверждение 2. Верхняя и нижняя границы диаграммы D_{TC} (множества соответственно максимальных и минимальных по взаимному вложению ее элементов) могут быть сформированы алгоритмом полиномиальной от размеров множеств A и ПУ вычислительной сложности.

Таким образом, порождая для всех текущих D_{TC} эти границы и далее в каждом конкретном случае направленным образом достраивая релевантные цепочки частичного порядка в D_{TC}, можно управлять перебором вариантов при поиске признаков вредоносной инсайдерской активности. При этом за счет оптимизации перебора и, если необходимо, за счет подключения дополнительных вычислительных ресурсов обеспечивается достижение ограничений процессно-реального времени.

Интерактивные сервисы, обеспечивающие сотрудникам службы безопасности оперативный доступ к деталям описаний диаграммы D_{TC}, вместе с возможностями проактивного отслеживания потенциально опасных направлений развития текущей ситуации оказываются дополнительным фактором повышения полноты и точности идентификации признаков вредоносной инсайдерской активности. При этом оперативный доступ к результатам аналитической обработки ранее выявленных признаков дает дополнительные инструменты управления качеством идентификации и противодействия потенциально опасной активности пользователей.

7 Заключение

Исследована задача поиска релевантных для дальнейшего анализа данных в Big Data, постоянно пополняемых новой информацией. В условиях жестких ограничений на время АД и ППР, т. е. процессно-реального времени, необходимо использовать средства ИАД. При этом существенные объемы анализируемых данных (эффект Big) и их постоянное пополнение новыми элементами (эф-

фект Open) — не единственные критически значимые факторы такого анализа.

В исследованных ситуациях достаточно часто встречались доказуемо трудноразрешимые комбинаторные проблемы, попадающие в те или иные известные классы вычислительной сложности. Для их решения использовались специальные проблемно-ориентированные частные решения. В связи с этим возникает необходимость предварительного выделения данных, релевантных целям поиска.

Как следствие, возникает дополнительная проблема «балансировки» детальности представления знаний и управления объемами необходимых для их обработки вычислений. Кроме того, эффективность исследований зависит от разработки надежных и быстрых программных сервисов для обеспечения эффективного интерактивного режима взаимодействия эксперта-аналитика и компьютерной системы ИАД. Соответствующая система АД и ППР должна предоставлять гибкие возможности направленной обработки знаний эксперта в том или ином формализованном виде. В частности, система должна позволять детализировать представление знаний о тех или иных аспектах анализируемой проблемы, «сужая» область исследования до конкретного «сектора» и сохраняя при этом «согласованность» с требованиями режима процессно-реального времени.

Следует отдельно подчеркнуть, что ответственность за результаты АД и ППР, выполненного с помощью компьютерной системы ИАД, ложится все-таки на эксперта-аналитика. Именно по этой причине сервисы интерактивного взаимодействия эксперта и системы ИАД, возможности реализации тех или иных эвристик в процессе поиска, сервисы управления перебором вариантов и т. п. оказываются не только полезными «инструментами» повышения полноты и точности поиска результатов, но и помогают эксперту сохранять понимание способа порождения и неформальную объяснимость получаемых заключений и рекомендаций.

Экспериментальное подтверждение работоспособности и результативности предлагаемой методики, а также ее процедурной реализации при решении задачи идентификации признаков вредоносной инсайдерской активности было получено в крупном отечественном коммерческом банке.

Литература

1. The flexible AI-powered Search & Discovery platform. https://www.algolia.com.
2. IBM Watson Discovery. https://www.ibm.com/cloud/watson-discovery.
3. Power your website with the world's best search. https://www.yext.com.
4. A powerful search experience for your website — without the learning curve. https://swiftype.com.
5. SearchUnify wins two silver Stevies — one in collaboration with Bluebeam — in 2021 Stevie Awards for Sales & Customer Service. https://www.searchunify.com.
6. ELASTIC: Search more, spend less. https://www.elastic.co.
7. Solr is the popular, blazing-fast, open source enterprise search platform built on Apache Lucene^{TM}. https://lucene.apache.org/solr.
8. Introduction to search with SPHINX. http://sphinxsearch.com.
9. Корпоративный поиск «Спутник». https://www.sputnik.ru/searchbox.
10. Архитектура платформы 1С-Предприятие: глобальный поиск. https://v8.1c.ru/platforma/globalnyy-poisk.
11. *Кон П. М.* Универсальная алгебра / Пер. с англ. — М.: Мир, 1968. 359 с. (*Cohn P. M.* Universal algebra. — New York, NY, USA: Harper and Row, 1965. 333 p.)
12. *Забежайло М. И.* О некоторых оценках сложности вычислений в ДСМ-рассуждениях // Искусственный интеллект и принятие решений, 2015. Часть I: № 1. С. 3–17; Часть II: № 2. С. 3–17.
13. *Грушо А. А., Забежайло М. И., Зацаринный А. А., Тимонина Е. Е.* О некоторых возможностях управления ресурсами при организации проактивного противодействия компьютерным атакам // Информатика и её применения, 2018. Т. 12. Вып. 1. С. 62–70.
14. *Simon J.* On the difference between one and many // Automata, languages and programming / Eds. A. Salomaa, M. Steinby. — Lecture notes in computer science ser. — Springer, 1977. Vol. 52. P. 480–491.
15. *Valiant L. G.* The complexity of enumeration and reliability problems // SIAM J. Comput., 1979. Vol. 8. Iss. 1. P. 410–421.
16. *Valiant L. G.* The complexity of computing the permanent // Theor. Comput. Sci., 1979. Vol. 8. P. 189–201.

Поступила в редакцию 04.04.2021

INTELLIGENT ANALYSIS OF BIG DATA EXTENDIBLE COLLECTIONS UNDER THE LIMITS OF PROCESS-REAL TIME

A. A. Grusho[1], M. I. Zabezhailo[2], D. V. Smirnov[3], and E. E. Timonina[1]

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation
[2]A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation
[3]Sberbank of Russia, 19 Vavilov Str., Moscow 117999, Russian Federation

Abstract: The problem how to extract relevant to the fixed goal data from regularly extended by new information collections of Big Data not braking given limits for data analysis and decision making (being in agreement with so-called process-real time restrictions) is discussed. The proposed approach is based on implementation of modern artificial intelligence techniques including knowledge representation and reasoning formalization for so-called Intelligent Data Analysis (IDA) computer systems. Some critical barriers preventing efficient application of this type IDA (e. g., computational complexity of some related to IDA combinatorial problems, including provable getting some of them in well-known classes of computationally hard problems, some characteristic features of knowledge representation and search iteration enumeration control, optimization of accuracy, and completeness of search results) are analyzed. A formalized description for the designed IDA set of procedures is presented. The discussed approach is illustrated by examples of its implementation in a corporate computer system of malicious insider activities identification and counteraction operating in a large Russian commercial bank.

Keywords: Big Data; process-real time; intelligent data analysis; information security; insider malicious activities

DOI: 10.14357/19922264210206

Acknowledgments

The paper was partially supported by the Russian Foundation for Basic Research (project 18-29-03081).

References

1. The flexible AI-powered Search & Discovery platform. Available at: https://www.algolia.com (accessed May 12, 2021).
2. IBM Watson Discovery. Available at: https://www.ibm.com/cloud/watson-discovery (accessed May 12, 2021).
3. Power your website with the world's best search. Available at: https://www.yext.com (accessed May 12, 2021).
4. A powerful search experience for your website — without the learning curve. Available at: https://swiftype.com (accessed May 12, 2021).
5. SearchUnify wins two silver Stevies — one in collaboration with Bluebeam — in 2021 Stevie Awards for Sales & Customer Service. Available at: https://www.searchunify.com (accessed May 12, 2021).
6. ELASTIC: Search more, spend less. Available at: https://www.elastic.co (accessed May 12, 2021).
7. Solr is the popular, blazing-fast, open source enterprise search platform built on Apache Lucene™. Available at: https://lucene.apache.org/solr/ (accessed May 12, 2021).
8. Introduction to Search with SPHINX. Available at: http://sphinxsearch.com (accessed May 12, 2021).
9. Korporativnyy poisk "Sputnik" [Corporate Search "Sputnik"]. Available at: https://www.sputnik.ru/searchbox (accessed May 12, 2021).
10. Arkhitektura platformy 1S-Predpriyatie: global'nyy poisk [1C-Enterprise Platform Architecture: Global search]. Available at: https://v8.1c.ru/platforma/globalnyy-poisk/ (accessed May 12, 2021).
11. Cohn, P. M. 1965. *Universal algebra*. New York, NY: Harper and Row. 333 p.
12. Zabezhailo, M. I. 2015. O nekotorykh otsenkakh slozhnosti vichisleniy v DSM-rassuzhdeniyakh [To the computational complexity of hypotheses generation in JSM-method]. *Iskusstvennyy intellect i prinyatie resheniy* [Artificial Intelligence and Decision Making]. Part I. 1:3–17; Part II. 2:3–17.
13. Grusho, A. A., M. I. Zabezhailo, A. A. Zatsarinny, and E. E. Timonina. 2018. O nekotorykh vozmozhnostyakh upravleniya resursami pri organizatsii proaktivnogo protivodeystviya komp'yuternym atakam [On some possibilities of resource management for organizing active counteraction to computer attacks]. *Informatika i ee Primeneniya — Inform. Appl.* 12(1):62–70.

14. Simon, J. 1977. On the difference between one and many. *Automata, languages and programming*. Eds. A. Salomaa and M. Steinby. Lecture notes in computer science ser. Berlin–Heidelberg: Springer. 52:480–491.
15. Valiant, L. G. 1979. The complexity of enumeration and reliability problems. *SIAM J. Comput.* 8:410–421.
16. Valiant, L. G. 1979. The complexity of computing the permanent. *Theor. Comput. Sci.* 8:189–201.

Received April 4, 2021

Contributors

Grusho Alexander A. (b. 1946) — Doctor of Science in physics and mathematics, professor, principal scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; grusho@yandex.ru

Zabezhailo Michael I. (b. 1956) — Doctor of Science in physics and mathematics, principal scientist, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; m.zabezhailo@yandex.ru

Smirnov Dmitry V. (b. 1984) — business partner for IT security department, Sberbank of Russia, 19 Vavilov Str., Moscow 117999, Russian Federation; dvlsmirnov@sberbank.ru

Timonina Elena E. (b. 1952) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119133, Russian Federation; eltimon@yandex.ru

НЕКОТОРЫЕ СВОЙСТВА СМЕСЕЙ НОРМАЛЬНЫХ РАСПРЕДЕЛЕНИЙ И ИХ ПРИЛОЖЕНИЯ К ЗАДАЧАМ МАГНИТОЭНЦЕФАЛОГРАФИИ*

М. Б. Гончаренко[1], Т. В. Захарова[2]

Аннотация: Рассматриваются различные свойства общих смесей вероятностных распределений. Особое внимание уделено случаю, когда смешиваемое распределение является нормальным. Установлены сходства в поведении нормальных смесей и нормальных распределений при трансформациях. Рассмотрено приложение к задачам исследования головного мозга методом магнитоэнцефалографии (МЭГ). Определены условия, при которых применима оценка Эйткена (обобщенного метода наименьших квадратов) для поиска источников нейрофизиологической активности в случае, когда распределение шума является нормальной смесью общего вида.

Ключевые слова: смеси распределений; смеси нормальных распределений; смеси распределений Стьюдента; смеси логнормальных распределений; смеси гамма-распределений; магнитоэнцефалография; МЭГ; обратная задача МЭГ; оценка Эйткена

DOI: 10.14357/19922264210207

1 Введение

Смеси вероятностных распределений общего вида (compound probability distribution) возникают в широком классе математических моделей, где параметры вероятностных распределений сами являются случайными величинами. Например, такая ситуация естественным образом возникает в процедуре байесовского вывода при подсчете апостериорного распределения. Другое распространенное приложение смесей — моделирование распределений с тяжелыми хвостами. Оно оказывается полезным для описания данных эксперимента с более высокой наблюдаемой дисперсией, чем предполагала оригинальная модель. Стоит отметить, что распределение случайных сумм также имеет вид смеси, а важный частный случай — конечной смеси распределений — широко используется при обработке неоднородных данных и, в частности, в задачах классификации наблюдений.

Данная статья посвящена исследованию свойств смесей нормальных законов. Повышенное внимание к нормальному распределению вызвано его широкой распространенностью в прикладных моделях анализа данных. Подробнее о нормальных смесях и их различных применениях можно прочесть в книгах [1, 2]. Знание рассмотренных в статье свойств поможет понять, как изменяются свойства модели при замене предположения о нормальности распределения какого-либо параметра (например, аддитивного шума) на смесь нормальных распределений. В рамках данной работы были обобщены результаты, полученные для конечных нормальных смесей в статье [3], на случай нормальных смесей общего вида. Отдельно рассматривается применение нормальных смесей в обратной задаче нейровизуализации (исследования распределения источников активности внутри головного мозга) методом МЭГ.

2 Базовые понятия

Чтобы исследовать свойства смесей нормальных распределений, сначала надо ввести строгое определение смеси вероятностных распределений. По этой теме имеется обширная литература (см. например [4, 5]), но в более современном виде понятие смеси дается в книге В. Ю. Королева [6], которое и будет процитировано ниже.

Рассмотрим функцию $F(x, \mathbf{y})$, определенную на множестве $\mathbb{R} \times \mathbb{Y}$.

Пусть $\mathbb{Y}$ — это некоторое подмножество m-мерного евклидова пространства, $\mathbb{Y} \subseteq \mathbb{R}^m$ при некотором $m \geqslant 1$, причем множество $\mathbb{Y}$ снабжено борелевской σ-алгеброй $\mathcal{B}$. Более того, при каждом

*Работа выполнена при частичной поддержке РФФИ (проект 19-07-00352) и в соответствии с программой Московского центра фундаментальной и прикладной математики.

[1]АО Интел, goncharenko.mir@yandex.ru

[2]Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, tvzaharova@mail.ru

фиксированном $\mathbf{y}$ функция $F(x, \mathbf{y})$ является функцией распределения по x, а при каждом фиксированном x функция $F(x, \mathbf{y})$ измерима по $\mathbf{y}$, т. е. для любых $x \in \mathbb{R}$ и $c \in \mathbb{R}$ выполнено условие $\{\mathbf{y} : F(x, \mathbf{y}) < c\} \in \mathcal{B}$. Пусть Q — вероятностная мера, определенная на измеримом пространстве $(\mathbb{Y}, \mathcal{B})$.

Функция распределения

$$H(x) = \int_{\mathbb{Y}} F(x, \mathbf{y}) \mathsf{Q}(d\mathbf{y}), \quad x \in \mathbb{R},$$

называется смесью функции распределения $F(x, \boldsymbol{y})$ ***по*** $\boldsymbol{y}$ ***относительно*** Q. Распределение $F(x, \mathbf{y})$ называется смешиваемым, в то время как мера Q задает смешивающее распределение.

Введем m-мерную случайную величину $\mathbf{Y}$: $\mathbf{Y}(y) \equiv y$, $y \in \mathbb{Y}$, определенную на вероятностном пространстве $(\mathbb{Y}, \mathcal{B}, \mathsf{Q})$. Тогда функция распределения $H(x)$ может быть записана в виде:

$$H(x) = \mathsf{E} F(x, \mathbf{Y}), \quad x \in \mathbb{R}.$$

Если $f(x, \mathbf{y})$ — плотность распределения, соответствующая функции распределения $F(x, \mathbf{y})$,

$$f(x, \mathbf{y}) = \frac{d}{dx} F(x, \mathbf{y}),$$

то смеси $H(x)$ соответствует плотность

$$h(x) = \mathsf{E} f(x, \mathbf{Y}) = \int_{\mathbb{Y}} f(x, \mathbf{y}) \mathsf{Q}(d\mathbf{y}), \quad x \in \mathbb{R}.$$

Далее будет рассмотрен важный частный случай вероятностных смесей: так называемая сдвиг/масштабная смесь. Введем определение согласно [6].

Пусть в определении, сформулированном выше, $m = 2$. Предположим, что вектор $\mathbf{y}$ имеет вид:

$$\mathbf{y} = (u, v),$$

где $u > 0$ и $v \in \mathbb{R}$, так что функция распределения $F(x, \mathbf{y})$ допускает представление

$$F(x, \mathbf{y}) = F\left(\frac{x - v}{u}\right), \quad x \in \mathbb{R}.$$

Тогда $\mathbb{Y}$ — это положительная полуплоскость, т. е. $\mathbb{Y} = \mathbb{R}^+ \times \mathbb{R}$, и функция распределения

$$H(x) = \int_{\mathbb{Y}} F\left(\frac{x - v}{u}\right) \mathsf{Q}(du, dv), \quad x \in \mathbb{R},$$

называется сдвиг/масштабной смесью функции распределения F ***относительно меры*** Q с параметром масштаба u и параметром сдвига (положения) v.

Если функция распределения F имеет плотность f, то смеси $H(x)$ соответствует плотность

$$h(x) = \int_{\mathbb{Y}} \frac{1}{u} f\left(\frac{x - v}{u}\right) \mathsf{Q}(du, dv), \quad x \in \mathbb{R}.$$

3 Основные результаты

3.1 Свойства нормальных смесей

Определение 1. Распределение случайной величины ξ является масштабной смесью нормальных распределений, если плотность $p_\xi(x)$ представима в виде:

$$p_\xi(x) = \int_0^\infty \frac{1}{\sigma} \varphi\left(\frac{x}{\sigma}\right) \mathsf{Q}(d\sigma), \quad x \in \mathbb{R},$$

где $\varphi(x)$ — плотность стандартного нормального распределения.

Далее исследуем свойства этих смесей и покажем, какие из свойств нормального распределения остаются справедливыми, а какие — нет.

Для полноты изложения приведем доказательство известного утверждения.

Утверждение 1. *Если плотность случайной величины ξ является масштабной смесью нормальных распределений, случайная величина X^2 имеет распределение $\chi^2(n)$ — хи-квадрат с n степенями свободы, X^2 и ξ независимы, то случайная величина $t = \xi / \sqrt{X^2/n}$ имеет плотность распределения, являющуюся смесью распределений Стьюдента с n степенями свободы.*

Доказательство. Рассмотрим плотность ξ, она имеет вид:

$$p_\xi(x) = \int_0^\infty \frac{1}{\sigma} \varphi\left(\frac{x}{\sigma}\right) \mathsf{Q}(d\sigma),$$

где $\varphi(x)$ — плотность стандартного нормального распределения.

Введем вспомогательную величину $\eta = \sqrt{X^2/n}$ с плотностью

$$p_\eta(y) = \frac{2ny(1/2)^{n/2}(ny^2)^{n/2-1}}{\Gamma(n/2)\, e^{ny^2/2}}, \text{ где } y \geq 0.$$

Тогда плотность частного ξ/η имеет вид:

$$p_{\xi/\eta}(z) = p_t(z) =$$
$$= \int_0^\infty y p_\eta(y) \left(\int_0^\infty \frac{1}{\sigma} \varphi\left(\frac{z\,y}{\sigma}\right) \mathsf{Q}(d\sigma) \right) dy =$$
$$= \int_0^\infty \frac{n^{n/2}\, y^n\, e^{-ny^2/2}}{2^{n/2-1}\Gamma(n/2)} \left(\int_0^\infty \frac{1}{\sigma} \varphi\left(\frac{z\,y}{\sigma}\right) \mathsf{Q}(d\sigma) \right) dy.$$

В силу теоремы Фубини изменив порядок интегрирования, проведем следующие преобразования:

$$p_t(z) = \int_0^\infty \frac{1}{\sqrt{2\pi}\,\sigma}\,\frac{n^{n/2}}{2^{n/2-1}\Gamma\,(n/2)}\times$$
$$\times\int_0^\infty y^n e^{-(1/2)\left(z^2/\sigma^2+n\right)y^2}dy\,\mathsf{Q}(d\sigma) =$$
$$=\int_0^\infty \frac{1}{\sqrt{2\pi}\,\sigma}\,\frac{n^{n/2}}{2^{n/2-1}\Gamma\,(n/2)}\times$$
$$\times\int_0^\infty (y^2)^{(n+1)/2-1}e^{-(1/2)\left(z^2/\sigma^2+n\right)y^2}\,d\left(\frac{y^2}{2}\right)\mathsf{Q}(d\sigma) =$$
$$=\int_0^\infty \frac{1}{\sqrt{2\pi}\,\sigma}\,\frac{n^{n/2}}{2^{n/2-1}\Gamma\,(n/2)}\,\frac{1}{2^{(1-n)/2}}\times$$
$$\times\left(\frac{z^2}{\sigma^2}+n\right)^{-(n+1)/2}\Gamma\left(\frac{n+1}{2}\right)\mathsf{Q}(d\sigma).$$

После упрощения подынтегрального выражения плотность $p_t(z)$ примет вид:

$$p_t(z) = \int_0^\infty \frac{1}{\sqrt{\pi}\,\sigma}\,\frac{\Gamma\left((n+1)/2\right)}{\Gamma\,(n/2)}\,\frac{1}{\sqrt{n}}\times$$
$$\times\left(\frac{n}{z^2/\sigma^2+n}\right)^{(n+1)/2}\mathsf{Q}(d\sigma) =$$
$$=\int_0^\infty \frac{1}{\sigma}\,\frac{1}{\sqrt{\pi\,n}}\,\frac{\Gamma\left((n+1)/2\right)}{\Gamma\,(n/2)}\times$$
$$\times\left(\frac{1}{1+z^2/(\sigma^2 n)}\right)^{(n+1)/2}\mathsf{Q}(d\sigma).$$

Используя обозначение $s_n(x)$ для плотности распределения Стьюдента с n степенями свободы

$$s_n(x) = \frac{\Gamma\left((n+1)/2\right)}{\Gamma(n/2)\sqrt{\pi n}}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2},\quad x\in\mathbb{R},$$

получим следующее выражение для плотности p_t:

$$p_t(x) = \int_0^\infty \frac{1}{\sigma}\,s_n\left(\frac{x}{\sigma}\right)\mathsf{Q}(d\sigma).$$

Таким образом, плотность p_t является масштабной смесью распределений Стьюдента относительно Q.

Утверждение доказано. □

Теорема 1. *Если случайная величина ξ имеет смешанное нормальное распределение относительно вероятностной меры Q, то случайная величина $\eta = \exp(\xi)$ имеет смешанное логнормальное распределение относительно Q. И наоборот, если η имеет смешанное логнормальное распределение, то $\xi = \ln\eta$ имеет смешанное нормальное распределение относительно одной и той же меры Q.*

Доказательство. Докажем сначала первое утверждение теоремы.

Получим выражение для плотности η в явном виде. По условию теоремы

$$\mathsf{P}(\xi\le x) = \int_0^\infty\int_{-\infty}^\infty \Phi\left(\frac{x-a}{\sigma}\right)\mathsf{Q}(d\sigma, da),$$

поэтому

$$F_\eta(x) = \mathsf{P}(\eta\le x)) = \mathsf{P}(\exp(\xi)\le x)) =$$
$$= \mathsf{P}(\xi\le \ln x)) = \int_0^\infty\int_{-\infty}^\infty \Phi\left(\frac{\ln x-a}{\sigma}\right)\mathsf{Q}(d\sigma, da),$$

где $\Phi(x)$ — функция распределения стандартного нормального закона, $x>0$.

Следовательно,

$$p_\eta(x) = \int_0^\infty\int_{-\infty}^\infty \frac{1}{x\sigma}\,\varphi\left(\frac{\ln x-a}{\sigma}\right)\mathsf{Q}(d\sigma, da).$$

Таким образом, плотность $p(x)$ смешиваемого распределения имеет вид

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}}\,e^{-(\ln x-a)^2/(2\sigma^2)},\quad x>0.$$

А значит, η имеет смешанное логнормальное распределение.

Для доказательства обратного утверждения теоремы воспользуемся следующей известной леммой.

Лемма 1. *Если функция $y=g(x)$ возрастает и дифференцируема, случайная величина ξ имеет плотность p_ξ, тогда плотность p_η случайной величины $\eta = g(\xi)$ определяется формулой*:

$$p_\eta(y) = p_\xi(g^{-1}(y))\,\frac{1}{g'(g^{-1}(y))}\,.$$

Итак, пусть плотность случайной величины ξ является смесью логнормальных распределений относительно Q. Тогда плотность случайной величины $\eta = \ln\xi$, с учетом леммы, равна

$$p_\eta(y) = p_\xi(\exp(y))\frac{1}{1/\exp(y)} = \exp(y) \times$$
$$\times \int_0^\infty \int_{-\infty}^\infty \frac{1}{\exp(y)\,\sigma}\,\varphi\left(\frac{\ln\exp(y) - a}{\sigma}\right)\mathsf{Q}(d\sigma, da) =$$
$$= \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma}\,\varphi\left(\frac{y-a}{\sigma}\right)\mathsf{Q}(d\sigma, da).$$

Таким образом, плотность случайной величины η является смесью нормальных распределений относительно той же смешивающей меры Q.

Теорема доказана. □

Замечание 1. Связь между нормальным и логнормальным распределениями сохраняется и для соответствующих смесей нормальных и логнормальных распределений. Данная теорема обобщает результаты, полученные авторами в статье [3].

Теорема 2. *Если плотность случайной величины ξ является масштабной смесью нормальных распределений, то случайная величина ξ^2 будет распределена с плотностью, являющейся масштабной смесью гамма-распределений, т. е.*

$$p_{\xi^2}(x) = \int_0^\infty \frac{1}{\sigma\sqrt{x}}\,\varphi\left(\frac{\sqrt{x}}{\sigma}\right)\mathsf{Q}(d\sigma) =$$
$$= \int_0^\infty \frac{1}{\sigma\sqrt{2\pi x}}\,e^{-x/(2\,\sigma^2)}\mathsf{Q}(d\sigma).$$

Доказательство. Выпишем функцию распределения для ξ^2 и проведем необходимые преобразования с использованием теоремы Фубини:

$$\mathrm{F}_{\xi^2}(x) = \mathsf{P}(\xi^2 \le x) = \mathsf{P}\left(-\sqrt{x} \le \xi \le \sqrt{x}\right) =$$
$$= \int_{-\sqrt{x}}^{\sqrt{x}} \int_0^\infty \frac{1}{\sigma}\,\varphi\left(\frac{x}{\sigma}\right)\mathsf{Q}(d\sigma)\,dx =$$
$$= \int_0^\infty \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sigma}\,\varphi\left(\frac{x}{\sigma}\right)dx\mathsf{Q}(d\sigma) =$$
$$= \int_0^\infty \left(\Phi\left(\frac{\sqrt{x}}{\sigma}\right) - \Phi\left(\frac{-\sqrt{x}}{\sigma}\right)\right)\mathsf{Q}(d\sigma) =$$
$$= \int_0^\infty \left(2\Phi\left(\frac{\sqrt{x}}{\sigma}\right) - 1\right)\mathsf{Q}(d\sigma) =$$
$$= 2\int_0^\infty \Phi\left(\frac{\sqrt{x}}{\sigma}\right)\mathsf{Q}(d\sigma) - 1.$$

Далее для нахождения плотности распределения ξ^2 продифференцируем функцию распределения, полученную выше:

$$p_{\xi^2}(x) = \frac{d}{dx}\,\mathrm{F}_{\xi^2}(x) = \int_0^\infty \frac{1}{\sigma\sqrt{x}}\,\varphi\left(\frac{\sqrt{x}}{\sigma}\right)\mathsf{Q}(d\sigma) =$$
$$= \int_0^\infty \frac{1}{\sigma\sqrt{2\pi\,x}}\,e^{-x/(2\sigma^2)}\mathsf{Q}(d\sigma).$$

Таким образом, плотность распределения случайной величины ξ^2 является масштабной смесью гамма-распределений.

Теорема доказана. □

Замечание 2. Эти три теоремы наглядно демонстрируют схожесть в поведении нормальных смесей и нормального распределения. Аналоги данных теорем для дискретных смесей можно найти в [3]. Но смеси нормальных распределений обладают особенностями поведения, отличающими их от нормального распределения. В частности, широко известный факт об эквивалентности свойств некоррелированности и независимости для компонент многомерного нормального распределения [7] уже не выполняется для конечной нормальной смеси, как показано в [3].

3.2 Обратная задача магнитоэнцефалографии

В статье [3] рассматривалось приложение конечных нормальных смесей для моделирования шума измерений активности головного мозга методом МЭГ. Магнитоэнцефалография — неинвазивная технология нейровизуализации, позволяющая исследовать электромагнитную активность человеческого мозга путем измерения магнитного поля непосредственно вблизи поверхности головы испытуемого (подробнее о МЭГ см. [8, 9]).

С помощью МЭГ можно исследовать различные аспекты функционирования головного мозга с высоким временны́м разрешением, сопоставимым со скоростью передачи нервного импульса (это качественно отличает МЭГ от другого популярного метода функциональной нейровизуализации — функциональной магнитно-резонансной томографии, фМРТ). Запись МЭГ представляет собой многоканальный сигнал, регистрируемый массивом сенсоров внутри специального шлема, под который помещается голова испытуемого. Отдельный интерес представляет локализация (указание точных координат и интенсивностей) источников

активности внутри головного мозга испытуемого. Для ее установления необходимо решить обратную задачу МЭГ.

Рассмотрим обратную задачу МЭГ:

$$Y = L\Theta + \mathcal{E}, \tag{1}$$

где $Y \in \mathbb{R}^n$ — измеряемые данные; $L \in \mathbb{R}^{n\times k}$ — оператор Био–Савара–Лапласа; $\Theta \in \mathbb{R}^k$ — неизвестные амплитуды источников; $\mathcal{E} \in \mathbb{R}^n$ — шум; k — количество источников активности; n — число МЭГ-сенсоров, $k \geq n$. Классический подход к решению подобных задач предполагает минимизацию нормы ошибки:

$$||\mathcal{E}||^2 = ||Y - L\Theta||^2 \to \min_{\Theta}.$$

О других подходах можно прочесть в статье [10]. Физические и математические свойства модели обратной задачи МЭГ рассматриваются в [11].

В теории линейной регрессии доказано, что у задачи (1) существует решение наименьших квадратов [12] при выполнении следующих условий:

- $\mathsf{E}\mathcal{E} = 0$ — математическое ожидание шума равно нулю;
- $\Sigma > 0$ — матрица ковариации ошибок положительно определена;
- $\operatorname{rank} L = n$, т. е. L — матрица полного строкового ранга.

Решение $\hat{\Theta}$, полученное методом взвешенных наименьших квадратов при выполнении обозначенных выше условий, называют оценкой Эйткена [12]:

$$\hat{\Theta} = \left(L^\top \Sigma^{-1} L\right)^{-1} L^\top \Sigma^{-1} Y. \tag{2}$$

Оценка (2) является несмещенной, состоятельной и оптимальной в классе всех линейных оценок [12].

Анализ реальных записей «пустой комнаты» (собственного шума сенсоров и окружающей среды, без испытуемого) показал, что зачастую распределение шума имеет сложную структуру с такими особенностями, как тяжелые хвосты, мультимодальность, несимметричность. Таким образом, для более адекватного описания реальных данных требуется более сложная модель шума.

В данном разделе будет рассматриваться модель шума, имеющего распределение в виде многомерной нормальной смеси общего вида с плотностью

$$h(\vec{x}) = \int_{\mathbb{Y}} f_y\left(\vec{x}; \vec{\mu_y}, \Sigma_y\right) \mathsf{Q}(dy), \tag{3}$$

где $f_y(\vec{x}; \vec{\mu_y}, \Sigma_y)$ — плотность k-мерного нормального распределения с вектором средних $\vec{\mu_y}$ и ковариационной матрицей Σ_y (для упрощения выкладок будем считать, что Σ_y невырождена $\forall y \in \mathbb{Y}$), $\vec{x} \in \mathbb{R}^k$. Далее будем использовать сокращенную запись $f_y(\vec{x})$.

В статье [3] была доказана теорема о том, что матрица ковариации конечной нормальной смеси положительно определена (в случае если в смеси есть компоненты с положительно определенными ковариационными матрицами).

Для доказательства соответствующей теоремы в случае общей нормальной смеси сначала докажем следующую лемму.

Лемма 2. *Матрица ковариации нормальной смеси* (3) *имеет вид*:

$$\Sigma = \mathsf{E}_y\Sigma_y + \mathsf{E}_y\left(\vec{\mu_y} - \mathsf{E}_y\vec{\mu_y}\right)\left(\vec{\mu_y} - \mathsf{E}_y\vec{\mu_y}\right)^\top,$$

где $\vec{\mu_y}$ и Σ_y — вектор средних и ковариационная матрица смешиваемого нормального распределения.

Доказательство. Пусть случайная величина $\vec{X}$ имеет смешанное k-мерное нормальное распределение с плотностью (3). Тогда ее матрица ковариации по определению равна

$$\Sigma = \mathsf{E}\vec{X}\vec{X}^\top - \mathsf{E}\vec{X}\mathsf{E}\vec{X}^\top.$$

Рассмотрим каждое из слагаемых подробнее:

$$\mathsf{E}\vec{X}\vec{X}^\top = \int_{\mathbb{R}^k} \vec{x}\vec{x}^\top \int_{\mathbb{Y}} f_y(\vec{x})\mathsf{Q}\,(dy)\,d\vec{x} =$$

$$= \int_{\mathbb{Y}}\int_{\mathbb{R}^k} \vec{x}\vec{x}^\top f_y(\vec{x})\,d\vec{x}\mathsf{Q}(dy) = \int_{\mathbb{Y}} \left(\Sigma_y + \vec{\mu_y}\vec{\mu_y}^\top\right)\mathsf{Q}(dy) =$$

$$= \mathsf{E}_y\left(\Sigma_y + \vec{\mu_y}\vec{\mu_y}^\top\right); \tag{4}$$

$$\mathsf{E}\vec{X} = \int_{\mathbb{R}^k} \vec{x} \int_{\mathbb{Y}} f_y(\vec{x})\mathsf{Q}(dy)\,d\vec{x} =$$

$$= \int_{\mathbb{Y}}\int_{\mathbb{R}^k} \vec{x} f_y(\vec{x})\,d\vec{x}\mathsf{Q}(dy) = \int_{\mathbb{Y}} \vec{\mu_y}\mathsf{Q}(dy) = \mathsf{E}_y\vec{\mu_y}. \tag{5}$$

Объединяя результаты (4) и (5) и перегруппировывая слагаемые, получим итоговое выражение для ковариационной матрицы в виде:

$$\Sigma = \mathsf{E}_y\Sigma_y + \mathsf{E}_y(\vec{\mu_y} - \mathsf{E}_y\vec{\mu_y})(\vec{\mu_y} - \mathsf{E}_y\vec{\mu_y})^\top. \tag{6}$$

Лемма доказана. □

Теорема 3. *Ковариационная матрица нормальной смеси* (3) *положительно определена*.

Доказательство. Рассмотрим случайную величину $\vec{X}$, имеющую смешанное нормальное распределение с плотностью (3). Ее ковариационная

матрица имеет вид (6). По определению матрица Σ является положительно определенной, если выполнено

$$u^\top \Sigma u > 0, \quad \forall\, u \in \mathbb{R}^k.$$

Распишем более подробно с учетом предыдущей леммы

$$\begin{aligned} &u^\top \Sigma u = \\ &= u^\top \mathsf{E}_y \Sigma_y u + u^\top \mathsf{E}_y (\vec{\mu_y} - \mathsf{E}_y \vec{\mu_y})(\vec{\mu_y} - \mathsf{E}_y \vec{\mu_y})^\top u = \\ &= \mathsf{E}_y u^\top \Sigma_y u + u^\top \mathsf{E}_y (\vec{\mu_y} - \mathsf{E}_y \vec{\mu_y})(\vec{\mu_y} - \mathsf{E}_y \vec{\mu_y})^\top u = \\ &= \mathsf{E}_y \underbrace{u^\top \Sigma_y u}_{>0} + \underbrace{u^\top \mathsf{E}_y (\vec{\mu_y} - \mathsf{E}_y \vec{\mu_y})(\vec{\mu_y} - \mathsf{E}_y \vec{\mu_y})^\top u}_{\geqslant 0}. \end{aligned}$$

Первое слагаемое строго положительно из-за положительной определенности матриц $\Sigma_y \forall y \in \mathbb{Y}$. Второе слагаемое есть ковариационная матрица $\vec{\mu_y}$; следовательно, она неотрицательно определена. В итоге получим, что

$$u^\top \Sigma u > 0.$$

Теорема доказана. □

Замечание 3. Теорема остается справедливой, даже если матрицы Σ_y вырождены при некотором $y \in A \subset \mathbb{Y}$. Это следует из того, что

$$\begin{aligned} u^\top \mathsf{E}_y \Sigma_y u &= u^\top \int\limits_{\mathbb{Y}} \Sigma_y \mathsf{Q}(dy) u = \\ &= u^\top \int\limits_{\mathbb{Y}\setminus A} \Sigma_y \mathsf{Q}(dy) u + u^\top \int\limits_{A} \Sigma_y \mathsf{Q}(dy) u \geqslant \\ &\geqslant u^\top \int\limits_{\mathbb{Y}\setminus A} \Sigma_y \mathsf{Q}(dy) u > 0, \ \forall u \in \mathbb{R}^k. \end{aligned}$$

Также из доказательства видно, что для справедливости теоремы достаточно, чтобы у смешиваемого распределения была положительно определенная матрица ковариации, а непосредственный вид смешиваемого распределения значения не имеет.

Таким образом, при использовании модели шума в виде нормальной смеси общего вида оценка интенсивностей источников с помощью обобщенного метода наименьших квадратов остается справедливой. Стоит отметить, что решение обратной задачи таким методом пользуется большой популярностью в прикладных нейрофизиологических исследованиях.

4 Заключение

В статье представлены базовые понятия непрерывных смесей вероятностных распределений и подробно рассмотрен частный случай нормальных смесей общего вида, определены законы распределения случайных величин, являющихся функциональным преобразованием случайных величин с плотностью в виде нормальной смеси общего вида.

Смеси распределений общего вида возникают в множестве прикладных задач, а также они используются как средство представления не-нормальных распределений. Для важного частного случая, где смешиваемое распределение является нормальным, были рассмотрены распределения трансформации смеси и установлены сходства в поведении нормальных смесей и нормального распределения. Также было доказано, что обобщенный метод наименьших квадратов поиска псевдообратного оператора остается применимым и в случае шума, имеющего распределение в виде смеси общего вида. Этот результат говорит о применимости широко распространенных методов решения обратной задачи МЭГ и в случае не-нормального шума, который может быть представлен в виде нормальной смеси. Такая ситуация часто встречается при обработке данных реальных экспериментов.

Литература

1. *Titterington D. M., Smith A. F. M., Makov U. E.* Statistical analysis of finite mixture distributions. — New York, NY, USA: Wiley, 1985. 243 p.
2. *McLachlan G. J., Peel D.* Finite mixture models. — New York, NY, USA: Wiley & Sons, 2000. 419 p.
3. *Гончаренко М. Б., Захарова Т. В.* Особенности поведения конечных смесей нормальных распределений // Вестник Московского университета. Сер. 15: Вычислительная математика и кибернетика, 2018. № 3. С. 30–36.
4. *Teicher H.* On the mixture of distributions // Ann. Math. Stat., 1960. Vol. 31. No. 1. P. 55–73.
5. *Teicher H.* Identifiability of finite mixtures // Ann. Math. Stat., 1963. Vol. 34. No. 4. P. 1265–1269.
6. *Королев В. Ю.* EM-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений: Теоретический обзор. — М.: ИПИ РАН, 2007. 94 с.
7. *Боровков А. А.* Математическая статистика. — 4-е изд. — М.: Лань, 2010. 704 с.
8. *Hamalainen M., Hari R., Ilmoniemi R. J., Knuutila J., Lounasmaa O. V.* Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain // Rev. Mod. Phys., 1993. Vol. 65. No. 2. P. 413–497.
9. *Захарова Т. В., Никифоров С. Ю., Гончаренко М. Б., Драницына М. А., Климов Г. А., Хазиахметов М. Ш., Чаянов Н. В.* Методы обработки сигналов для лока-

лизации невосполнимых областей головного мозга // Системы и средства информатики, 2012. Т. 22. № 2. С. 157–175.

10. *Гончаренко М. Б., Захарова Т. В.* Вероятностный подход к решению обратной задачи МЭГ // Системы и средства информатики, 2018. Т. 28. № 1. С. 35–52.
11. *Zakharova T. V., Karpov P. I., Bugaevskii V. M.* Localization of the activity source in the inverse problem of magnetoencephalography // Comput. Math. Model., 2017. Vol. 28. No. 2. P. 148–157.
12. *Дрейпер Н. Р., Смит Г.* Прикладной регрессионный анализ / Пер. с англ. — 3-е изд. — М.: Диалектика, 2016. 912 с. (*Draper N., Smith H.* Applied regression analysis. — 3rd ed. — New York, NY, USA: John Wiley & Sons, Inc., 1998. 736 p.)

Поступила в редакцию 27.09.2020

SOME PROPERTIES OF GAUSSIAN MIXTURES AND APPLICATIONS TO MAGNETOENCEPHALOGRAPHY PROBLEMS

M. B. Goncharenko[1] and T. V. Zakharova[2,3]

[1]INTEL A/O, 17-4 Krylatskaya Str., Moscow 121614, Russian Federation
[2]Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation
[3]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The article is dedicated to research of various properties of compound probability distributions (mixture distributions). Special attention is paid to the case when the mixed distribution is Gaussian. The authors establish the similarities in the behavior of Gaussian mixtures and Gaussian distributions during transformations. The authors study applications to magnetoencephalographic brain research. The authors determine the conditions under which the Aitken estimator (generalized least squares) is applicable for localization of sources of neurophysiologic activity in the case of noise having compound Gaussian distribution.

Keywords: compound distributions; compound Gaussian distribution; compound Student distribution; compound lognormal distribution; compound gamma distributions; magnetoencephalography; MEG; inverse MEG problem; Aitken's estimator

DOI: 10.14357/19922264210207

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research (project 19-07-00352). The research was conducted in accordance with the program of the Moscow Center for Fundamental and Applied Mathematics.

References

1. Titterington, D. M., A. F. M. Smith, and U. E. Makov. 1985. *Statistical analysis of finite mixture distributions*. New York, NY: Wiley. 243 p.
2. McLachlan, G., and D. Peel. 2000. *Finite mixture models*. New York, NY: Wiley & Sons. 419 p.
3. Goncharenko, M. B., and T. V. Zakharova. 2018. Osobennosti povedeniya konechnykh smesey normal'nykh raspredeleniy [Features of behavior of finite mixtures of normal distributions]. *Vestnik Moskovskogo Universiteta. Ser. 15: Vychislitel'naya matematika i kibernetika* [Bull. Moscow State University. Ser. 15: Comput. Math., Cybern.] 3:30–36.
4. Teicher, H. 1960. On the mixture of distributions. *Ann. Math. Stat.* 31(1):55–73.
5. Teicher, H. 1963. Identifiability of finite mixtures. *Ann. Math. Stat.* 34(4):1265–1269.
6. Korolev, V. Yu. 2007. *EM-algoritm, ego modifikatsii i ikh primenenie k zadache razdeleniya smesey veroyatnostnykh raspredeleniy. Teoreticheskiy obzor* [EM algorithm modifications and their application to the separation of mixtures of probability distributions. Theoretical review]. Moscow: IPI RAN. 94 p.
7. Borovkov, A. A. 2010. *Matematicheskaya statistika* [Mathematical statistics]. 4th ed. Moscow: Lan. 704 p.
8. Hamalainen, M., R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. 1993. Magnetoencephalography — theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.* 65(2):413–497.

9. Zakharova, T. V., S. Yu. Nikiforov, M. B. Goncharenko, M. A. Dranitsyna, G. A. Klimov, M. Sh. Khaziakhmetov, and N. V. Chayanov. 2012. Metody obrabotki signalov dlya lokalizatsii nevospolnimykh oblastey golovnogo mozga [Signal processing methods for localization of nonrenewable brain regions]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 22(2):157–175.
10. Goncharenko, M. B., and T. V. Zakharova. 2018. Veroyatnostnyy podkhod k resheniyu obratnoy zadachi magnitoentselografii [Probabilistic approach to solving the magnetoencephalography inverse problem]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 28(1):35–52.
11. Zakharova, T. V., P. I. Karpov, and V. M. Bugaevskii. 2017. Localization of the activity source in the inverse problem of magnetoencephalography. *Comput. Math. Model.* 28(2):148–157.
12. Draper, N., and H. Smith. 1998. *Applied regression analysis*. 3rd ed. New York, NY: John Wiley & Sons, Inc. 736 p.

Received September 27, 2020

Contributors

Goncharenko Miroslav B. (b. 1991) — software development engineer for graphics, INTEL A/O, 17-4 Krylatskaya Str., Moscow 121614, Russian Federation; goncharenko.mir@yandex.ru

Zakharova Tatiana V. (b. 1962) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; lsa@cs.msu.ru

МЯГКИЕ ВЫЧИСЛЕНИЯ В ЗАДАЧАХ МЕДИЦИНСКОЙ ДИАГНОСТИКИ

М. П. Кривенко[1]

Аннотация: В последние годы возрастает значение информатики для интерпретации и анализа данных с использованием вычислительных методов, в частности так называемых «мягких» вычислений (Soft Computing, SC). Рассматриваются возможности применения SC для решения проблем, связанных с медициной, и в особенности в задачах поддержки принятия решений. При этом демонстрируется, что не следует искусственно задействовать новации, тем более что ценой небольших усилий можно обратиться к классическим подходам, методологически строгим и приводящим к гарантированным результатам. Несомненный интерес к методологиям SC в различных дисциплинах (генетика, физиология, радиология, кардиология, неврология и т. д.) показывает, что их изучение крайне плодотворно, и ожидается, что будущие исследования в медицине будут использовать соответствующие методы в большей степени, чем сегодня, и для решения более сложных задач.

Ключевые слова: медицина; мягкие вычисления; референсные значения; байесовский подход

DOI: 10.14357/19922264210208

1 Введение

«Мягкие» вычисления — не новый термин, он часто применяется в компьютерных науках и информационных технологиях. Инструментарий «мягких» вычислений использует технику нечетких систем (нечеткие множества, нечеткая логика, «грубые» множества), искусственные нейронные сети, генетические алгоритмы и эволюционное моделирование, в том числе иммунные алгоритмы, алгоритмы роевого интеллекта. Приведенное описание состава «мягких» вычислений не является единственным, для этого достаточно сравнить аннотации для SC на сайтах трех издательств: Elsevier, Springer и Wiley.

У каждой компоненты SC есть свои достоинства. В сочетании они представляют собой не просто набор инструментов, а скорее партнерство, в котором каждый предлагает свою методологию для решения общей проблемы. Главное, что выделяется в настоящее время при описании сути SC, — это единство отдельных подходов (методологий), которые работают синергетически и предоставляют в той или иной форме гибкие возможности оценки неоднозначных ситуаций в реальной жизни. Цель применения SC состоит в том, чтобы учесть допуски, неточности, неопределенности данных, приблизительность аргументации и правдоподобия при построении вывода для получения гибких, но надежных недорогих решений.

В последние годы наблюдается рост биоинформатики и медицинской информатики с использованием вычислительных методов для интерпретации и анализа данных. Если ограничиться базовыми методологиями нечеткой логики (Fuzzy Logic, FL), нейронных сетей (Neural Networks, NN) и применения генетического алгоритма (Genetic Algorithm, GA), то поиск в базе данных Medline по названиям работ за два последних десятилетия даст результаты, представленные в табл. 1. Из нее видно: число публикаций во второй декаде увеличилось на 130%, что действительно свидетельствует о возрастании в медицине внимания к SC; доля смешанных методов по отношению к базовым (чистым) подходам не изменилась и составляет 1%, т. е. о всеохватывающем синергетическом эффекте SC пока говорить не приходится; бросающееся в глаза внимание к генетическим алгоритмам демонстрирует проявление близости гуманитарной природы медицины и сути эволюционных алгоритмов.

Цель данной статьи состоит в том, чтобы продемонстрировать возможности применения SC для решения проблем, связанных с медициной, и в особенности в задачах поддержки принятия решений. Несомненный интерес к изучению методологий SC в различных дисциплинах (генетика, физиология, радиология, кардиология, неврология и т. д.) показывает, что их освоение очень плодотворно, и ожидается, что будущие исследования в медицине будут использовать соответствующие методы в большей

[1] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, mkrivenko@ipiran.ru

Таблица 1 Число упомянутых в Medline работ по группам методов за два последних десятилетия

Группа методов	Применяемые методы	2001–2010 гг.	2011–2020 гг.
Одиночные	Только FL	206	332
	Только NN	4 633	11 431
	Только GA	677	970
Смешанные	Только FL и NN	2	8
	Только FL и GA	2	1
	Только NN и GA	66	119
	FL и NN и GA	1	1
Общая	Всевозможные комбинации FL, NN, GA	5 587	12 862

степени, чем сегодня, и для решения более сложных задач. При этом не следует искусственно задействовать новации, тем более что ценой небольших усилий можно обратиться к классическим подходам, методологически строгим и приводящим к гарантированным результатам.

2 Области применения

В медицине можно выделить четыре области: фундаментальная, диагностическая, клиническая и хирургическая.

Фундаментальная медицина характеризуется наличием множества явлений с крайне сложным взаимодействием отдельных элементов, с одной стороны, и небогатым опытом моделирования подобных явлений, с другой. По этой причине она подходит для всех методологий SC. В первую очередь речь идет о биохимических, генетических, физиологических и фармакологических отраслях. Не следует забывать о междисциплинарных направлениях типа науки о здоровье.

Результаты исследования источников по использованию методологий FL–NN приводят к таким отраслям, как цитология, физиология, генетика и биостатистика. Ярким проявлением заинтересованности фундаментальной науки в SC служит обработка изображений [1].

Другой интересный пример — применение FL–NN в генетике. В частности, комбинированный метод для прогнозирования раковой ткани по данным экспрессии генов [2] — нейрокомпьютинг, основанный на знаниях, — использовался для создания нечетких правил, которые указывают на гены, тесно связанные с определенными типами рака.

Категория FL–GA представлена несколькими областями исследований, причем в первую очередь речь идет о генетике. В исследовании экспрессии генов [3] предпочтение было отдано FL-контроллеру, настроенному с помощью GA. Был предложен интерпретируемый классификатор с точной и компактной базой нечетких правил для анализа данных на микрочипах.

Категория NN–GA стала наиболее предпочтительной для дисциплин фундаментальных наук: биохимии, биостатистики, генетики, гистологии, патологии, фармакологии и физиологии.

Обращает внимание пример использования указанной методологии в фармакологии при изучении взаимодействия лекарственных средств с живыми организмами. В [4] исследована модель для прогнозирования проницаемости структурно различных препаратов в зависимости от выбранных молекулярных дескрипторов с использованием искусственных нейронных сетей, а генетический алгоритм используется для выбора подмножества дескрипторов, которые наилучшим образом описывают степень проникновения препарата.

Не всегда привлечение отдельных методов SC оказывается ожидаемым: авторы [5] использовали методологию GA в качестве алгоритма обучения при сравнении производительности NN, они отметили, что GA оказывается неэффективным при тонкой настройке локального поиска.

Диагностическая медицина в основном включает в себя радиологические и клинические лабораторные исследования.

Преимущественно SC применяются в интервенционной радиологии. При этом превалируют FL–NN приложения.

Сегментация цифровых изображений — один из наиболее важных этапов их анализа. Изображения всегда содержат значительный уровень шума (вызванного действиями оператора, оборудованием и окружающей средой), что может привести к серьезным неточностям при сегментации. Авторы [6] эффективно использовали метод FL–NN в своих исследованиях для решения задач магнитно-резонансной томографии.

Клиническая медицина стала наиболее популярной и подходящей областью для применения методологии SC. Сравнение предпочтений в этой

области указывает на первенство FL–NN и NN–GA, при этом на первых ролях такие отрасли, как кардиология, неврология, терапия критических состояний, анестезиология, физическая медицина и реабилитационная медицина.

В области анестезии есть ряд убедительных примеров использования адаптивных систем для контроля артериального давления, обезболивания, паралича, потери сознания и септического шока. Примером последних публикаций на эту тему может служить [7], где исследуется применимость адаптивной нейро-нечеткой стратегии в управлении анестезией с обратной связью. Показывается, что построенный контроллер адаптивен и устойчив к проблеме вариабельности между пациентами и по отношению к каждому из них, а также эффективен в условиях наличия шума измерительных устройств.

Большая часть работы по использованию адаптивных систем в области кардиологии была направлена на кардиостимуляторы. Здесь активно используется модель ANFIS (Adaptive Neuro-Fuzzy Inference System), исследование [8] которой продемонстрировало, что ее показатели точности выше, чем у автономной модели NN.

По сравнению с FL–NN методология NN–GA оказывается менее предпочтительной в клинической науке. Несмотря на это, имеют место некоторые успешные применения в нефрологии, пульмонологии, неврологии, психиатрии, физической и реабилитационной медицине.

В [9] представлен способ классификации результатов аускультации легких: GA использован с целью поиска оптимальной структуры и параметров обучения NN для лучшего прогнозирования динамики шумов.

Хирургия. В [10] указывается, что результаты поиска в базе данных Medline показали, что хирургическая наука не обращается к тем или иным элементам SC. Причина проста: существует общее мнение о хирургии, что она в основном связана с навыками хирурга. А работы, освещающие «околохирургические» проблемы, относят обычно к другим отраслям медицины. Примером может служить [11], где предлагаются решения на основе SC для прогнозирования послеоперационной выживаемости при раке легкого. Встречаются работы по использованию SC при описании объектов хирургической практики: в исследованиях [12] применяется подход, основанный на эволюционной технике вкупе с искусственной нейронной сетью для нахождения решений нелинейных обыкновенных дифференциальных уравнений второго порядка, используемых в качестве модели роговицы глаза.

В последнее время безразличие хирургии к SC уходит в прошлое. Современные операции чаще всего проводятся с использованием роботизированных хирургических инструментов и другого оборудования, поэтому все более актуальными становятся работы типа определения инструментально-тканевого нажима в роботизированной лапароскопической хирургии с использованием нейро-эволюционных нечетких систем [13].

Другие примеры применения SC в различных отраслях медицины можно найти в [10], а именно: критическая медицина (управление искусственной вентиляцией легких, контроль развития септического шока во время пребывания пациентов в отделении интенсивной терапии, обнаружение нормальной и искаженной плетизмограммы); неврология (распознавание стадии сна на основе знаний и обработки биосигналов); физическая и реабилитационная медицина (контроль интенсивности физической нагрузки спортсмена, диагностика и контроль тремора); дерматология (дифференциальная диагностика эритематозно-сквамозных очагов); эндокринология (моделирование динамики сложных метаболических систем применительно к внутриклеточной кинетике тиамина); онкология (поддержка принятия решений по идентификации подтипа клеток острого лимфобластного лейкоза у детей с использованием данных по экспрессии генов; прогнозирование рака молочной железы и простаты); гастроэнтерологии (дифференциальная диагностика при сложных желудочно-кишечных расстройствах, таких как диспепсический синдром или хронический панкреатит, когда симптомы у пациента могут лежать на пересечении группы расстройств).

3 Поддержка принятия решений на основе референсных значений

Поддержка принятия решений при диагностировании заболеваний по результатам клинических исследований может рассматриваться как задача классификации набора показателей, входящих в анализы отдельного пациента, где класс — это определенный синдром. При этом результат классификации должен представлять собой совокупность диагнозов, упорядоченную в направлении от наиболее к наименее правдоподобным. При нечетком задании исходных данных совершенно естественно обратиться к возможностям SC, тем более что для этого имеются соответствующие наработки в виде монографий (например, [14]) и отдельных публикаций.

Таблица 2 Характеристики синдрома «первичный гиперпаратиреоз»

№	f_i	Критическая область показателя S_i	Вес показателя ν_i
1	Кальций в крови	$> 2{,}55$	8
2	Паратгормон в крови	> 55	7
...	...	...	...
13	Отношение клиренс кальция / клиренс креатинина	$> 0{,}02$	8
	Сумма весов		87

В случае когда показатели суть категориальные и/или числовые, нечеткие рассуждения можно перевести на формальный язык математической статистики и теории принятия решений. Это позволит опереться на строгую методологическую основу, формулировать и решать возникающие задачи, ясно отдавать себе отчет в том, что и как получено в результате.

Далее будем считать, что при построении классификатора заданы следующие характеристики: частота встречаемости ситуаций типа «определенный диапазон значений показателя – конкретный синдром», причем такие данные имеются далеко не по всем комбинациям «показатель–синдром»; частота встречаемости в генеральной совокупности данных тех или иных синдромов.

Показатели в зависимости от сложности их задания разбиваются на следующие группы (примеры в скобках приведены для стандартного клинического анализа крови без указания единиц измерения): бинарные (например, наличие или отсутствие теста по определению общих триглицеридов); однопороговые (например, при определении АСТ норма составляет диапазон от 0 до 40); многопороговые (например, при определении ЛПНП-бета оптимальные значения суть 0,3–2,4, близкие к оптимальным — от 2,5 до 3,1, пограничо высокие — от 3,2 до 3,9, высокие — от 4,0 до 4,8, очень высокие — более 4,8).

Для конкретности далее рассмотрим задачу, которая возникла в ходе разработки информационно-аналитической автоматизированной системы «Мегалит» [15].

Распространенным на практике алгоритмом классификации набора показателей служит метод голосования, когда для показателей определены области нормальных/аномальных референсных значений, а вес синдрома определяется через долю случаев, когда значение показателя отклонилось от нормы.

Пусть показатель, входящий в состав анализа, обозначен как f_i, $i = 1, \ldots, d$, а некоторый синдром — D_j, $j = 1, \ldots, M$. Конкретно речь шла о 33 показателях (в частности, кальций в крови, паратгормон в крови и т. п.), $d = 33$, и 6 синдромах (в частности, первичный гиперпаратиреоз, вторичный гиперпаратиреоз и т. п.), $M = 6$. Представление об исходных данных для одного из синдромов можно получить из табл. 2 (используется нотация специалистов предметной области, для сокращения записи опущены единицы измерения). Вес показателя, приведенный в последнем столбце этой таблицы, взят из опыта диагностирования и фактически означает относительную частоту встречаемости. Заметим, что в матрице «показатель–синдром» всего 198 элементов, при этом 129 комбинаций оказались незаданными.

Таким образом, алгоритм классификации методом голосования, обрабатывающий данные для комбинаций f_i, D_j и формирующий в результате перечень весов диагнозов, принимает вид: для $j = 1, \ldots, M$ подсчитать веса диагнозов по формуле

$$W_j = \frac{\sum_{i=1}^{d} \mathbf{1}_{S_i}(f_i)\nu_i}{\sum_{i=1}^{d} \nu_i},$$

где $\mathbf{1}_A(x)$ — индикатор множества A, после чего следует отдать предпочтение диагнозу с наибольшим значением веса. Данный алгоритм является эвристическим и по виду соответствует результату обучения NN.

Использование референсных интервалов для принятия решений создает предпосылки для привлечения нечетких множеств, а фактическое отсутствие обоснованных правил постановки диагноза по совокупности показателей — для мобилизации методов SC, в частности нейронных сетей. В последнем случае естественней опереться на обучающую выборку, которой в описываемой ситуации нет. Поэтому была сделана попытка продвинуться по пути формирования вероятностной модели так, чтобы можно было моделировать исходные данные, а также ставить и решать задачу построения решающих правил. Ее основу составило представление отдельного показателя в виде категориальной пе-

Таблица 3 Представление многопорогового показателя

j	S_2	ν_2	Вероятности для пяти категорий				
1	> 55	7	0,3/2	0,3/2	0,7/3	0,7/3	0,7/3
2	> 200	9	0,1/3	0,1/3	0,1/3	0,9/2	0,9/2
3	> 500	9	0,1/4	0,1/4	0,1/4	0,1/4	0,9/1
4	> 10	5	0,5/1	0,5/4	0,5/4	0,5/4	0,5/4
5	> 55	8	0,2/2	0,2/2	0,8/3	0,8/3	0,8/3
6	N/A	N/A	0,2	0,2	0,2	0,2	0,2

ременной с вероятностями появления отдельных значений.

Примером однопорогового показателя служит «кальций в крови». Соответствующая переменная принимает два значения: категория «1» (значения $f_1 < 2{,}55$); категория «2» (значения $f_1 \geq 2{,}55$). Вероятности появления категорий зависят от класса — синдрома: так, для $j = 1$ имеем $S_1 = (> 2{,}55)$, $\nu_1 = 8$ и вероятности категорий $(0{,}2; 0{,}8)$, а для $j = 2$ имеем $S_1 = (< 2{,}55)$, $\nu_1 = 8$ и вероятности категорий $(0{,}8; 0{,}2)$. При описании случаев, когда значения показателей не заданы, все исходы принимались равновероятными. Подобных однопороговых показателей среди исходных оказалось всего 23.

В качестве примера многопорогового показателя рассмотрим «паратгормон в крови», для которого большинству синдромов соответствуют свои пороговые значения (см. столбец S_2 табл. 3), поэтому число категорий возрастает до пяти. Соответствующие вероятности для категорий приведены в последних столбцах табл. 3. Запись вероятности в виде, например, 0,7/3 наглядно демонстрирует использование столбца весов ν_2. Для обозначения ситуации, когда значение показателя не задано, используется аббревиатура N/A. Подобных многопороговых показателей набралось 8.

Примером бинарных показателей может служить «вид мочевой инфекции», сводящийся к фиксации наличия или отсутствия мочевиноразлагающей флоры. В этом случае вводятся две категории: если появление показателя не задано, то вероятности категорий $(0{,}5; 0{,}5)$, если же мочевиноразлагающая флора встречается, вероятности категорий $(1{,}0; 0{,}0)$. Подобные двоичные показатели встретились 2 раза.

Перечисленные группы показателей исчерпывают все имеющиеся на данный момент варианты. Таким образом, результаты анализов описываются случайным вектором, каждая координата которого имеет дискретное распределение, в принципе, с разным числом исходов. Теперь появляется возможность строить оптимальные правила классификации, в частности на байесовских принципах.

Сравнительный анализ классификаторов. Построенная вероятностная модель данных позволяет провести сравнительный анализ алгоритмов классификации как методом голосования, так и байесовским. В качестве критерия использовалась вероятность правильной классификации p_{CC}. Для некоторого классификатора, заданного с помощью вероятностных характеристик, можно построить матрицу результатов классификации $\mathbf{C}$, где c_{ij} — вероятность отнесения классифицируемого элемента из i-го класса к j-му. Тогда $p_{\mathrm{CC}} = \mathrm{tr}\,(\mathbf{C})$.

Для реализации байесовского классификатора необходимо определиться с его параметрами: в качестве функции потерь рассматривалась единичная, объединение отдельных показателей осуществлялось в предположении их независимости; априорные вероятности классов π_j принимались либо равными (случай отсутствия информации о вероятностях появления синдромов), либо «реальными», полученными на основе анализа прецедентов; в последнем случае речь шла о значениях превалентности $\boldsymbol{\pi} = (0{,}33; 0{,}49; 0{,}05; 0{,}02; 0{,}08; 0{,}03)$.

Дискретный характер данных создает иллюзию, что матрицу можно посчитать точно, но из-за проблем многомерности реализовать это практически не представляется возможным: общее число различных значений классифицируемых элементов составляет 733 835 427 840 ($\sim 10^{12}$). Поэтому был применен метод моделирования, в основе которого лежит генератор из смеси дискретных распределений с вероятностями появления элементов смеси $\boldsymbol{\pi}$.

Результаты сравнительного анализа двух классификаторов при числе экспериментов, равном 10^4, сведены в табл. 4. Большое число экспериментов позволяет уверенно говорить о явных преимуществах байесовского классификатора.

Как кажется на первый взгляд, условия задачи поддержки принятия решений при постановке диагноза предопределяют использование SC. Но дополнительные усилия по формированию вероятностной модели принесли свои плоды: методо-

Таблица 4 Результаты сравнительного анализа классификаторов на основе оценок p_{CC}

Способ классификации	Равновероятные синдромы	«Реальные» вероятности синдромов
Байесовский классификатор	74,3%	82,4%
Классификация методом голосования	62,6%	59,7%

логически выверенное построение модели помогло формализовать постановку задачи анализа; использование байесовского классификатора обеспечило учет априорной информации в бóльшем объеме (например, вероятности появления синдромов), гарантировало наименьшие потери, а в случае единичной функции потерь и наименьшие значения вероятности ошибки; предложенная вероятностная модель данных позволила на основе имеющейся информации проводить исследования методов и алгоритмов анализа данных и систематизировать алгоритмы принятия решений, строить схемы последующего обучения процедур обработки.

Понятно, что все это удается сделать для оказавшейся приемлемой сложности реальной задачи. Если это не так, то фактор многомерности данных (многочисленность показателей, синдромов, отдельных значений показателей) просто сделает актуальной постановку задач о снижении размерности, приведет к необходимости развития соответствующих методов.

4 Заключение

Анализ применения SC проведен по основным медицинским приложениям. На основании этого можно предварительно прогнозировать дальнейшее развитие этой технологии в медицине, получить представление о возникающих типовых задачах анализа данных. Из представленного описания различных приложений можно увидеть, что методы SC применяются в широком спектре областей медицины для визуализации, диагностики, прогнозирования и контроля протекающих процессов.

Методологии SC, которые имитируют человеческий стиль мышления и принятия решений при решении сложных проблем, может преодолеть недостатки традиционных систем поддержки принятия медицинских решений, основанных на статистических моделях и традиционных методах искусственного интеллекта. Как и все другие приближенные методы, SC-методы имеют относительные преимущества и недостатки.

Примеры постановки рассмотренных задач с использованием «мягких» вычислений свидетельствуют о возможности применения соответствующих методов, а не о необходимости прибегать к ним. Единственным исключением служит ситуация, когда исходная проблема формулируется на языке лингвистических переменных. Но и здесь не очевидно, что надо обращаться к нечеткому выводу, нейронным сетям, генетическим алгоритмам и т. п., а не пытаться воспользоваться методологически выверенным математическим подходом, дающим четкое представление о возникающих ограничениях и гарантирующим свойства построенных решений.

Надо признать, что при использовании «мягких» вычислений возникают множества скрытых параметров, задача выбора значений которых не решается, а подменяется общими рекомендациями о влиянии этих параметров на итоговое качество принимаемых решений.

Крайне затрудняет освоение полученных результатов то, что источники по «мягким» вычислениям подчас содержат, казалось бы, мелкие неточности, но порождающие большие сомнения. Публикации по «мягким» вычислениям подчас носят рекламный характер, они хороши для знакомства с новыми идеями в области создания и применения информационных технологий, но оказываются бесполезными, а иногда и вредными, с точки зрения специалистов, которым необходимо решать актуальные наукоемкие задачи практики.

Все перечисленные недостатки не снижают интерес к описанному подходу, а скорее подогревают. Причина этого в первую очередь в том, что далеко не все задачи практики удается пока формально полностью поставить и тем более решить. На данный момент самым важным является улучшение понимания сильных и слабых сторон идей и методов SC, использование их лучших возможностей.

Литература

1. *Zanaty E. A., Ghoniemy S.* Medical image segmentation techniques: An overview // Int. J. Informatics Medical Data Processing, 2016. Vol. 1. No. 1. P. 16–37.
2. *Catto J. W. F., Linkens D. A., Abbod M. F., Chen M., Burton J. L., Feeley K. M., Hamdy F. C.* Artificial intelligence

in predicting bladder cancer outcome: A comparison of neuro-fuzzy modeling and artificial neural networks // Clin. Cancer Res., 2003. Vol. 9. No. 11. P. 4172–4177.

3. *Ho S. Y., Hsieh C. H., Chen H. M., Huang H. L.* Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis // Biosystems, 2006. Vol. 85. P. 165–176.
4. *Agatonovic-Kustrin S., Evans A., Alany R. G.* Prediction of corneal permeability using artificial neural networks // Pharmazie, 2003. Vol. 58. No. 10. P. 725–729.
5. *Ghaffari A., Abdollahi H., Khoshayand M. R., Bozchaloi S., Dadgar A., Rafiee-Tehrani M.* Performance comparison of neural network training algorithms in modeling of bimodal drug delivery // Int. J. Pharm., 2006. No. 327. P. 126–138.
6. *Shen S., Sandham W., Granat M., Sterr A.* MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural-network optimization // IEEE T. Inf. Technol. B., 2005. Vol. 9. No. 3. P. 459–467.
7. *Li R., Wu1 Q., Liu J., Wu Q., Li C., Zhao Q.* Monitoring depth of anesthesia based on hybrid features and recurrent neural network // Front. Neurosci. — Switz., 2020. Vol. 14. Art. 26.
8. *Ubeyli E. D., Guler I.* Adaptive neuro-fuzzy inference systems for analysis of internal carotid arterial Doppler signals // Comput. Biol. Med., 2005. Vol. 35. No. 8. P. 687–702.
9. *Guler I., Polat H., Ergun U.* Combining neural network and genetic algorithm for prediction of lung sounds // J. Med. Syst., 2005. Vol. 29. No. 3. P. 217–231.
10. *Yardimci A.* Soft computing in medicine // Appl. Soft Comput., 2009. Vol. 9. P. 1029–1043.
11. *Iraji M. S.* Prediction of post-operative survival expectancy in thoracic lung cancer surgery with soft computing // J. Appl. Biomed., 2017. Vol. 15. Iss. 2. P. 151–159.
12. *Waseem W., Sulaiman M., Alhindi A., Alhakami H.* Soft computing approach based on fractional order DPSO algorithm designed to solve the corneal model for eye surgery // IEEE Access, 2020. Vol. 8. P. 61576–61592.
13. *Mozaffari A., Behzadipour S., Kohani M.* Identifying the tool-tissue force in robotic laparoscopic surgery using neuro-evolutionary fuzzy systems and a synchronous self-learning hyper level supervisor // Appl. Soft Comput., 2014. Vol. 14. Part A. P. 12–30.
14. *Chang M.* Artificial intelligence for drug development, precision medicine, and healthcare. — Boca Raton, FL, USA: Chapman & Hall/CRC, 2020. 355 p.
15. *Голованов С. А., Кривенко М. П., Савченко П. А., Сивков А. В., Сучков А. П.* Информационно-аналитическая автоматизированная система «Мегалит» в оптимизации диагностики и лечения мочекаменной болезни // Информатика и её применения, 2013. Т. 7. Вып. 4. С. 82–93.

Поступила в редакцию 01.12.2020

SOFT COMPUTING IN PROBLEMS OF MEDICAL DIAGNOSTICS

M. P. Krivenko

Institute of Informatics Problems, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: In recent years, the importance of informatics has increased for the interpretation and analysis of data using computational methods, in particular, the so-called “soft” computing (Soft Computing — SC). The article discusses the possibilities of using SC for solving problems related to medicine and, especially, problems of decision support. At the same time, it is demonstrated that one should not artificially use innovations, especially since, at the cost of little effort, one can turn to classical approaches that are methodologically rigorous and lead to guaranteed results. The undoubted interest in the study of SC methodologies in various disciplines (genetics, physiology, radiology, cardiology, neurology, etc.) demonstrates that their study is extremely fruitful and it is expected that future research in medicine will use the corresponding methods to a greater extent than today and for more complex tasks.

Keywords: medicine; soft computing; reference values; Bayesian approach

DOI: 10.14357/19922264210208

References

1. Zanaty, E. A., and S. Ghoniemy. 2016. Medical image segmentation techniques: An overview. *Int. J. Informatics Medical Data Processing* 1(1):16–37.
2. Catto, J. W. F., D. A. Linkens, M. F. Abbod, M. Chen, J. L. Burton, K. M. Feeley, and F. C. Hamdy. 2003. Artificial intelligence in predicting bladder cancer outcome: A comparison of neuro-fuzzy modeling and artificial neural networks. *Clin. Cancer Res.* 9(11):4172–4177.
3. Ho, S. Y., C. H. Hsieh, H. M. Chen, and H. L. Huang. 2006. Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. *Biosystems* 85:165–176.

4. Agatonovic-Kustrin, S., A. Evans, and R. G. Alany. 2003. Prediction of corneal permeability using artificial neural networks. *Pharmazie* 58(10):725–729.
5. Ghaffari, A., H. Abdollahi, M. R. Khoshayand, S. Bozchaloi., A. Dadgar, and M. Rafiee-Tehrani. 2006. Performance comparison of neural network training algorithms in modeling of bimodal drug delivery. *Int. J. Pharm.* 327:126–138.
6. Shen, S., W. Sandham, M. Grana, and A. Sterr. 2005. MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural-network optimization. *IEEE T. Inf. Technol. B.* 9(3):459–467.
7. Li, R., Q. Wu1, J. Liu, Q. Wu, C. Li, and Q. Zhao. 2020. Monitoring depth of anesthesia based on hybrid features and recurrent neural network. *Front. Neurosci. —Switz.* 14:26. 11 p.
8. Ubeyli, E. D., and I. Guler. 2005. Adaptive neuro-fuzzy inference systems for analysis of internal carotid arterial Doppler signals. *Comput. Biol. Med.* 35(8):687–702.
9. Guler, I., H. Polat H., and U. Ergun. 2005. Combining neural network and genetic algorithm for prediction of lung sounds. *J. Med. Syst.* 29(3):217–231.
10. Yardimci, A. 2009. Soft computing in medicine. *Appl. Soft Comput.* 9:1029–1043.
11. Iraji, M. S. 2017. Prediction of post-operative survival expectancy in thoracic lung cancer surgery with soft computing. *J. Appl. Biomed.* 15(2):151–159.
12. Waseem, W., M. Sulaiman, A. Alhindi, and H. Alhakami. 2020. Soft computing approach based on fractional order DPSO algorithm designed to solve the corneal model for eye surgery. *IEEE Access* 8:61576–61592.
13. Mozaffari, A., S. Behzadipour, and M. Kohani. 2014. Identifying the tool-tissue force in robotic laparoscopic surgery using neuro-evolutionary fuzzy systems and a synchronous self-learning hyper level supervisor. *Appl. Soft Comput.* 14(A):12–30.
14. Chang, M. 2020. *Artificial intelligence for drug development, precision medicine, and healthcare*. Boca Raton, FL, USA: Chapman & Hall/CRC. 355 p.
15. Golovanov, S. A., M. P. Krivenko, P. A. Savchenko, A. V. Sivkov, and A. P. Suchkov. 2013. Informatsionno-analiticheskaya avtomatizirovannaya sistema "Megalit" v optimizatsii diagnostiki i lecheniya mochekamennoy bolezni [The information-analytical computer system "Megalith" in optimization of the diagnosis and treatment of urolithiasis]. *Informatika i ee Primeneniya — Inform. Appl.* 7(4):82–93.

Received December 1, 2020

Contributor

Krivenko Michail P. (b. 1946) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; mkrivenko@ipiran.ru

МЕТОД ВЫПРЯМЛЕНИЯ ИСКАЖЕННЫХ ИЗ-ЗА МУЛЬТИКОЛЛИНЕАРНОСТИ КОЭФФИЦИЕНТОВ В РЕГРЕССИОННЫХ МОДЕЛЯХ

М. П. Базилевский[1]

Аннотация: При построении регрессионной модели из-за сильной мультиколлинеарности объясняющих переменных происходит искажение ее коэффициентов, в частности их знаков, что негативно сказывается на интерпретационных качествах такой регрессии. Статья посвящена разработке метода выпрямления искаженных из-за мультиколлинеарности коэффициентов. В основе этого метода лежит свойство, которым обладают ранее предложенные автором модели полносвязной линейной регрессии. Исследована нелинейная система, по которой осуществляется оценивание полносвязных регрессий. Показано, что решение этой системы может быть получено численно с помощью метода простых итераций. Предложен способ выбора неизвестных лямбда-параметров в полносвязной регрессии. Установлено, что в многофакторных полносвязных моделях при сильной корреляции всех факторов знаки коэффициентов при переменных во вторичном уравнении совпадают с соответствующими знаками коэффициентов корреляции. Для выпрямления искаженных коэффициентов на основе проведенного исследования разработан алгоритм «Selection B». Разработанный метод выпрямления успешно продемонстрирован на примере моделирования валового внутреннего продукта (ВВП) России.

Ключевые слова: регрессионный анализ; модель полносвязной линейной регрессии; мультиколлинеарность; интерпретация; численный метод; ВВП России

DOI: 10.14357/19922264210209

1 Введение

При оценивании неизвестных параметров регрессионных моделей, например с помощью метода наименьших квадратов (МНК), на практике часто приходится сталкиваться с проблемой мультиколлинеарности [1, 2]. Это негативное явление возникает из-за наличия сильной корреляции между двумя или более независимыми переменными. Мультиколлинеарность факторов приводит к искажению коэффициентов в уравнении регрессии. В частности, их знаки могут противоречить теоретическим предпосылкам решаемой задачи. Поэтому построенная при мультиколлинеарности регрессионная модель остается годной в лучшем случае только для прогнозирования, но никак не для интерпретации и принятия каких-либо правильных управленческих решений.

Проблема мультиколлинеарности на сегодня еще окончательно не решена. Существуют лишь несколько основных подходов к ее устранению [3, 4].

Во-первых, это метод исключения [3]. Он заключается в том, что на основе матрицы парных коэффициентов корреляции определяются пары сильно коррелированных объясняющих переменных. Затем из каждой пары исключается тот фактор, который слабее коррелирует с зависимой переменной. После чего по оставшимся факторам оценивается регрессионная модель. Недостаток данного подхода состоит в том, что в полученном уравнении из-за исключения нельзя изучать совместное влияние всех исходных объясняющих переменных на объясняемую.

Во-вторых, метод главных компонент [5]. С помощью этого способа происходит формирование новых и не коррелирующих между собой переменных — главных компонент, являющихся линейными комбинациями старых переменных. К сожалению, в этом случае возникает проблема с интерпретацией главных компонент.

В-третьих, ридж-регрессия [6]. В этом случае в формулу для МНК-оценивания регрессии добавляется так называемый коэффициент регуляризации, который решает проблему мультиколлинеарности. Однако нет четких правил для выбора этого коэффициента. И нет гарантии, что в полученной модели коэффициенты будут удовлетворять содержательному смыслу задачи.

Как справедливо отмечено в работе [7], все эти методы ориентированы на устранение только вычислительных проблем. Но проблема, связанная с построением интерпретируемых при мультиколлинеарности регрессионных моделей, остается нерешенной.

[1]Иркутский государственный университет путей сообщения, кафедра математики, mik2178@yandex.ru

Целью данной работы ставилась разработка метода выпрямления искаженных из-за мультиколлинеарности коэффициентов линейных регрессионных моделей. Основой для этого метода послужило замеченное автором свойство двухфакторных полносвязных регрессий [8, 9], состоящее в том, что в их вторичных уравнениях знаки коэффициентов при объясняющих переменных совпадают с соответствующими знаками коэффициентов корреляции. Это же свойство может быть справедливо и для многофакторных моделей полносвязной линейной регрессии, впервые предложенных в работе [10].

2 Многофакторная модель полносвязной линейной регрессии без выходной переменной

Пусть x_{ij}, $i=\overline{1,n}$, $j=\overline{1,m}$, — наблюдаемые значения m входных переменных $x_1, x_2, \ldots, x_m$. Предположим, что существуют их «истинные» значения $x^*_{i1}, x^*_{i2}, \ldots, x^*_{im}$, $i=\overline{1,n}$, связанные с наблюдаемыми значениями соотношениями

$$x_{ij} = x^*_{ij} + \varepsilon_i^{(x_j)}\,,\ \ i=\overline{1,n}\,,\ j=\overline{1,m}\,. \qquad (1)$$

Предположим, что между «истинными» переменными $x^*_1, x^*_2, \ldots, x^*_m$ имеют место функциональные зависимости

$$x^*_j = a_j + b_j x^*_m\,,\ \ j=\overline{1,m-1}\,, \qquad (2)$$

где a_j и b_j, $j=\overline{1,m-1}$, — неизвестные параметры.

Совокупность уравнений (1) и (2) называется многофакторной моделью полносвязной линейной регрессии без выходной переменной [10]. Для ее оценивания применим взвешенный метод наименьших полных квадратов (ВМНПК):

$$\begin{aligned} S = \lambda_1 \sum_{i=1}^{n} (x_{i1} - a_1 - b_1 x^*_{im})^2 + \\ + \lambda_2 \sum_{i=1}^{n} (x_{i2} - a_2 - b_2 x^*_{im})^2 + \cdots \\ \cdots + \lambda_{m-1} \sum_{i=1}^{n} (x_{i,m-1} - a_{m-1} - b_{m-1} x^*_{im})^2 + \\ + \sum_{i=1}^{n} (x_{im} - x^*_{im})^2 \to \min\,, \end{aligned} \qquad (3)$$

где $\lambda_1, \lambda_2, \ldots, \lambda_{m-1}$ — положительные весовые коэффициенты (лямбда-параметры).

В работе [10] показано, что если лямбда-параметры известны, то решение задачи (3) осуществляется по следующему алгоритму.

1. Находятся оценки $\tilde b_1, \tilde b_2, \ldots, \tilde b_{m-1}$ параметров $b_1, b_2, \ldots, b_{m-1}$. Для этого численно решается нелинейная система вида:

$$\begin{aligned} b_p \Bigg(D_{x_m} + \sum_{j=1}^{m-1} \lambda_j^2 b_j^2 D_{x_j} + \\ + 2 \sum_{j_1=1}^{m-2} \sum_{j_2=j_1+1}^{m-1} \lambda_{j_1} \lambda_{j_2} b_{j_1} b_{j_2} K_{x_{j_1} x_{j_2}} + \\ + 2 \sum_{j=1}^{m-1} \lambda_j b_j K_{x_j x_m} \Bigg) = \Bigg(1 + \sum_{j=1}^{m-1} \lambda_j b_j^2 \Bigg) \times \\ \times \Bigg(\sum_{j=1}^{m-1} \lambda_j b_j K_{x_j x_p} + K_{x_m x_p} \Bigg)\,,\ \ p=\overline{1,m-1}\,. \end{aligned} \qquad (4)$$

2. Определяются оценки $\tilde a_1, \tilde a_2, \ldots, \tilde a_{m-1}$ параметров $a_1, a_2, \ldots, a_{m-1}$ по формулам:

$$\tilde a_j = \overline{x}_j - \tilde b_j \overline{x}_m\,,\ \ j=\overline{1,m-1}\,. \qquad (5)$$

3. Вычисляются оценки «истинных» значений переменной x_m по формулам:

$$\begin{aligned} \tilde x^*_{im} = \Bigg(1 + \sum_{j=1}^{m-1} \lambda_j \tilde b_j^2 \Bigg)^{-1} \Bigg(- \sum_{j=1}^{m-1} \lambda_j \tilde a_j \tilde b_j + \\ + \sum_{j=1}^{m-1} \lambda_j \tilde b_j x_{ij} + x_{im} \Bigg)\,,\ \ i=\overline{1,n}\,. \end{aligned} \qquad (6)$$

Очевидно, что если абсолютные значения парных коэффициентов корреляции переменных $x_1, x_2, \ldots, x_m$ равны 1, то при оценивании полносвязной регрессии по критерию (3) все остатки будут равны 0, а ее оценки $\tilde b_i$, $i=\overline{1,m-1}$, будут совпадать с МНК-оценками соответствующих парных регрессий. А знаки этих оценок согласуются со знаками соответствующих коэффициентов корреляции $r_{x_i x_m}$, $i=\overline{1,m-1}$, т. е. справедливы условия $\tilde b_i r_{x_i x_m} > 0$, $i=\overline{1,m-1}$. Значит, они будут справедливы и при сильной корреляции факторов.

3 Численный метод решения нелинейной системы (4)

Систему (4) нетрудно привести к виду:

$$H_p b_p^2 + B_p b_p + C_p = 0\,,\ \ p=\overline{1,m-1}\,, \qquad (7)$$

где

$$H_p = \lambda_p \left(K_{x_m x_p} + \sum_{j \in \{1,\ldots,m-1\} \setminus \{p\}} \lambda_j b_j K_{x_j x_p} \right);$$

$$B_p = D_{x_m} + \sum_{j \in \{1,\ldots,m-1\} \setminus \{p\}} \lambda_j^2 b_j^2 D_{x_j} +$$

$$+ 2 \sum_{j_1 \in \{1,\ldots,m-2\} \setminus \{p\}} \sum_{j_2 \in \{j_1+1,\ldots,m-1\} \setminus \{p\}} \lambda_{j_1} \lambda_{j_2} b_{j_1} b_{j_2} K_{x_{j_1} x_{j_2}} +$$

$$+ 2 \sum_{j \in \{1,\ldots,m-1\} \setminus \{p\}} \lambda_j b_j K_{x_j x_m} - \lambda_p D_{x_p} - \lambda_p D_{x_p} \times$$

$$\times \sum_{j \in \{1,\ldots,m-1\} \setminus \{p\}} \lambda_j b_j^2;$$

$$C_p = - \left(1 + \sum_{j \in \{1,\ldots,m-1\} \setminus \{p\}} \lambda_j b_j^2 \right) \times$$

$$\times \left(K_{x_m x_p} + \sum_{j \in \{1,\ldots,m-1\} \setminus \{p\}} \lambda_j b_j K_{x_j x_p} \right).$$

Тогда систему (7) можно представить в виде:

$$H_p \left(b_p - b_{p,1}^* \right) \left(b_p - b_{p,2}^* \right) = 0, \quad p = \overline{1, m-1}, \quad (8)$$

где $b_{p,1}^* = (-B_p - \sqrt{\mathrm{Disc}_p})/(2H_p)$ и $b_{p,2}^* = (-B_p + \sqrt{\mathrm{Disc}_p})/(2H_p)$ — корни p-го квадратного трехчлена системы (7); $\mathrm{Disc}_p = B_p^2 - 4H_pC_p$ — дискриминанты p-го квадратного трехчлена системы (7), которые, как видно, всегда положительны.

Понятно, что система (8) равносильна совокупности 2^{m-1} систем

$$\begin{cases} b_1 = b_{1,1}^*; \\ b_2 = b_{2,1}^*; \\ \ldots \\ b_{m-1} = b_{m-1,1}^*; \end{cases} \quad \begin{cases} b_1 = b_{1,2}^*; \\ b_2 = b_{2,1}^*; \\ \ldots \\ b_{m-1} = b_{m-1,1}^*; \end{cases} \quad \ldots$$

$$\ldots \quad \begin{cases} b_1 = b_{1,2}^*; \\ b_2 = b_{2,2}^*; \\ \ldots \\ b_{m-1} = b_{m-1,2}^*. \end{cases} \quad (9)$$

Покажем, что решение задачи (3) удовлетворяет только системе

$$\begin{cases} b_1 = b_{1,2}^*; \\ b_2 = b_{2,2}^*; \\ \ldots \\ b_{m-1} = b_{m-1,2}^*. \end{cases} \quad (10)$$

Вторые частные производные функции (3) имеют вид:

$$\frac{\partial^2 S}{\partial b_p^2} = 2\lambda_p n \left(1 + \sum_{j=1}^{m-1} \lambda_j b_j^2 \right)^{-2} \left[2H_p b_p + B_p - \right.$$

$$\left. - 2\frac{b_p}{n} \left(1 + \sum_{j=1}^{m-1} \lambda_j b_j^2 \right) \frac{\partial S}{\partial b_p} \right], \quad p = \overline{1, m-1}. \quad (11)$$

Для того чтобы функция (3) имела минимум в некоторой точке, матрица Гессе, составленная из частных производных второго порядка, должна быть положительно определенной. По критерию Сильвестра для положительно определенной матрицы Гессе все ее элементы на главной диагонали (11) должны быть положительными. А из 2^{m-1} систем (9) это условие выполняется только для случая (10). Поэтому для нахождения оценок полносвязной регрессии вместо системы (4) достаточно решить систему (10). Если $m \geq 3$, то для этого можно воспользоваться методом простых итераций. При $m = 3$ можно также применить метод подстановки.

Как уже отмечалось, до решения системы (4) необходимо задать значения лямбда-параметров. По мнению автора, рациональным будет выбор таких значений этих параметров, при которых суммарное аппроксимационное качество полносвязной регрессии (1), (2) будет наилучшим. Для этого введем аддитивный коэффициент детерминации

$$R_{\mathrm{add}}^2 = \sum_{j=1}^{m} R_{x_j}^2,$$

где $R_{x_j}^2$ — коэффициент детерминации для переменной x_j полносвязной регрессии (1), (2).

Сформулируем следующую оптимизационную задачу:

$$\sum_{j=1}^{m} R_{x_j}^2 \to \max,$$

которая, по определению $R_{x_j}^2$, равносильна задаче

$$\frac{\sum_{i=1}^{n} \left(\varepsilon_i^{(x_1)} \right)^2}{D_{x_1}} + \frac{\sum_{i=1}^{n} \left(\varepsilon_i^{(x_2)} \right)^2}{D_{x_2}} + \ldots$$

$$\ldots + \frac{\sum_{i=1}^{n} \left(\varepsilon_i^{(x_m)} \right)^2}{D_{x_m}} \to \min. \quad (12)$$

А задача (12) эквивалентна задаче (3) при $\lambda_1 = D_{x_m}/D_{x_1}, \lambda_2 = D_{x_m}/D_{x_2}, \ldots, \lambda_m = D_{x_m}/D_{x_{m-1}}$. Таким образом, для полученных значений лямбда-параметров аппроксимационное качество многофакторной полносвязной регрессии (1), (2) будет наилучшим.

4 Многофакторная модель полносвязной линейной регрессии с выходной переменной и алгоритм «Straight B»

Дополним набор входных переменных $x_1, x_2, \ldots, x_m$ выходной переменной y, которая сильно коррелирует с ними. Свяжем оцененные «истинные» значения, например, переменной $\tilde{x}^*_m$ со значениями переменной y моделью парной линейной регрессии:

$$y_i = c_0 + c_1 \tilde{x}^*_{im} + \varepsilon_i\,, \quad i = \overline{1,n}\,, \tag{13}$$

где c_0 и c_1 — неизвестные параметры, которые находятся с помощью обычного МНК.

Совокупность уравнений (1), (2), (13) называется многофакторной моделью полносвязной линейной регрессии с выходной переменной y [10]. Если параметры $\lambda_1, \lambda_2, \ldots, \lambda_{m-1}$ известны, то ее оценки находятся в два этапа.

1. С помощью МНПК оценивается полносвязная регрессия без выходной переменной (1), (2).
2. С помощью МНК оценивается модель парной линейной регрессии (13).

Пусть оцененная модель (1), (2) , (13) имеет вид:

$$\tilde{y} = \tilde{c}_0 + \tilde{c}_1 \tilde{x}^*_m\,; \tag{14}$$

$$\tilde{x}^*_j = \tilde{a}_j + \tilde{b}_j \tilde{x}^*_m\,, \quad j = \overline{1,m-1}\,;$$

$$\tilde{x}^*_m = A_0 + \sum_{j=1}^{m} A_j x_j\,, \tag{15}$$

где

$$A_0 = -\left(1 + \sum_{j=1}^{m-1} \lambda_j \tilde{b}_j^2\right)^{-1} \sum_{j=1}^{m-1} \lambda_j \tilde{a}_j \tilde{b}_j\,;$$

$$A_j = \lambda_j \tilde{b}_j \left(1 + \sum_{j=1}^{m-1} \lambda_j \tilde{b}_j^2\right)^{-1}, \quad j = \overline{1,m-1}\,;$$

$$A_m = \left(1 + \sum_{j=1}^{m-1} \lambda_j \tilde{b}_j^2\right)^{-1}.$$

Используя (15), перепишем уравнение (14) в виде:

$$\tilde{y} = \theta_0 + \sum_{j=1}^{m} \theta_j x_{ij}\,, \tag{16}$$

где $\theta_0 = \tilde{c}_0 + \tilde{c}_1 A_0$; $\theta_j = \tilde{c}_1 A_j$, $j = \overline{1,m}$.

Уравнение (16) называется вторичным уравнением многофакторной модели полносвязной линейной регрессии [10].

Как уже отмечалось, при сильной корреляции входных переменных $x_1, x_2, \ldots, x_m$ коэффициенты уравнения (15) удовлетворяют условиям

$$A_j r_{x_j x_m} > 0\,, \quad j = \overline{1,m-1}\,; \quad A_m > 0\,, \tag{17}$$

а при сильной корреляции y с этими переменными угловой коэффициент уравнения (14) — условию

$$\tilde{c}_1 r_{y x_m} > 0\,. \tag{18}$$

Перемножив неравенства (17) на (18), получим $\tilde{c}_1 A_j r_{x_j x_m} r_{y x_m} > 0$, $j = \overline{1,m-1}$, и $\tilde{c}_1 A_m r_{y x_m} > 0$. Отсюда, учитывая, что знаки произведений $r_{x_j x_m} r_{y x_m}$ совпадают со знаками $r_{y x_j}$, $j = \overline{1,m-1}$, следует, что

$$\theta_j r_{y x_j} > 0\,, \quad j = \overline{1,m}\,,$$

т. е. знаки коэффициентов при объясняющих переменных во вторичном уравнении (16) совпадают с соответствующими знаками коэффициентов корреляции $r_{y x_j}$, $j = \overline{1,m}$.

На основе проведенных автором исследований разработан следующий алгоритм «Straight B», реализующий метод выпрямления искаженных коэффициентов (МВИК) для многофакторных регрессионных моделей.

1. При $\lambda_1 = D_{x_m}/D_{x_1}, \lambda_2 = D_{x_m}/D_{x_2}, \ldots, \lambda_{m-1} = D_{x_m}/D_{x_{m-1}}$ из системы (10) численно находятся оценки $\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_{m-1}$, затем по формулам (5) — коэффициенты $\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_{m-1}$ и, наконец, по формулам (6) — оцененные «истинные» значения переменной $\tilde{x}^*_m$.
2. С помощью МНК оценивается модель (13).
3. Путем подстановки (15) в равенство (14) определяется искомое уравнение регрессии.

5 Моделирование валового внутреннего продукта России

Для демонстрации МВИК решалась задача моделирования ВВП России. Для этого были использованы статистические данные с сайта Федеральной службы государственной статистики (https://rosstat.gov.ru) за период с 2000 по 2018 гг. по следующим семи переменным: y — ВВП России (млрд руб.); x_1 — среднемесячная заработная плата в России (руб.); x_2 — численность безработных в России (тыс. чел.); x_3 — потребление

Матрица парных коэффициентов корреляции

	y	x_1	x_2	x_3	x_4	x_5	x_6
y	1	0,997	−0,843	0,953	0,985	0,942	0,997
x_1	0,997	1	−0,827	0,95	0,988	0,94	0,998
x_2	−0,843	−0,827	1	−0,888	−0,805	−0,922	−0,831
x_3	0,953	0,95	−0,888	1	0,92	0,981	0,954
x_4	0,985	0,988	−0,805	0,92	1	0,917	0,987
x_5	0,942	0,94	−0,922	0,981	0,917	1	0,937
x_6	0,997	0,998	−0,831	0,954	0,987	0,937	1

электроэнергии в России (млн кВт·ч); x_4 — продукция сельского хозяйства России (млрд руб.); x_5 — грузооборот железнодорожного транспорта России (млрд т·км); x_6 — оборот розничной торговли (млн руб.).

Матрица парных коэффициентов корреляции для этих переменных представлена в таблице.

Как видно из таблицы, все объясняющие переменные тесным образом коррелируют между собой. Следовательно, в оцененной по исходным данным модели множественной линейной регрессии будет присутствовать эффект мультиколлинеарности. С помощью МНК была построена следующая регрессионная модель:

$$\tilde{y} = 24236{,}7 + 1{,}399x_1 - 2{,}136x_2 - 0{,}0067x_3 - \\ - 0{,}225x_4 - 3{,}073x_5 + 0{,}0013x_6\,. \quad (19)$$

Коэффициент детерминации модели (19) $R^2 = 0{,}996$. Как и оказалось, из-за мультиколлинеарности знаки коэффициентов при переменных x_3, x_4 и x_5 не согласуются с экономическим смыслом задачи. Так, по регрессии (19) можно сделать абсурдный вывод о том, что для повышения ВВП России требуется снижать объемы производства продукции сельского хозяйства.

Для МВИК на основе алгоритма «Selection B» на языке программирования hansl эконометрического пакета Gretl была написана соответствующая программа. С помощью этой программы было получено вторичное уравнение полносвязной регрессии:

$$\tilde{y} = -54754{,}02 + 0{,}409x_1 - 4{,}765x_2 + 0{,}0693x_3 + \\ + 3{,}475x_4 + 15{,}73x_5 + 0{,}00054x_6\,. \quad (20)$$

Коэффициент детерминации модели (20) $R^2 = 0{,}971$. Как видно, теперь знаки абсолютно всех коэффициентов согласуются с экономическим смыслом задачи. При этом по отношению к модели (19) качество регрессии (20) снизилось незначительно, поэтому ее можно использовать не только для интерпретации, но и для прогнозирования.

Литература

1. *Gunst R. F., Webster J. T.* Regression analysis and problems of multicollinearity // Commun. Stat. Theory, 1975. Vol. 4. P. 277–292.
2. *Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T.* Best subset selection for eliminating multicollinearity // J. Oper. Res. Soc. Jpn., 2017. Vol. 60. No. 3. P. 321–336.
3. *Ферстер Э., Ренц Б.* Методы корреляционного и регрессионного анализа / Пер. с нем. — М.: Финансы и статистика, 1983. 304 с. (*Förster E., Rönz B.* Methoden der korrelation und regressionsanalyse. — Berlin: Verlag Die Wirtschaft, 1979. 369 p.)
4. *Chatterjee S., Hadi A. S.* Regression analysis by example. — 5th ed. — Hoboken, NJ, USA: Wiley, 2012. 424 p.
5. *Jolliffe I. T.* Principal component analysis. — New York, NY, USA: Springer-Verlag, 2002. 488 p.
6. *Hoerl A. E., Kennard R. W.* Ridge regression: Biased estimation for nonorthogonal problems // Technometrics, 1970. Vol. 12. P. 55–67.
7. *Мокшина С. И., Шуршикова Г. В., Щекунских С. С.* Метод построения содержательно интерпретируемых регрессионных моделей в условиях мультиколлинеарности // Современная экономика: проблемы и решения, 2017. № 5(89). С. 81–94.
8. *Базилевский М. П.* Синтез модели парной линейной регрессии и простейшей EIV-модели // Моделирование, оптимизация и информационные технологии, 2019. Т. 7. № 1(24). С. 170–182.
9. *Базилевский М. П.* Исследование двухфакторной модели полносвязной линейной регрессии // Моделирование, оптимизация и информационные технологии, 2019. Т. 7. № 2(25). С. 80–96.
10. *Базилевский М. П.* Многофакторные модели полносвязной линейной регрессии без ограничений на соотношения дисперсий ошибок переменных // Информатика и её применения, 2020. Т. 14. Вып. 2. С. 92–97.

Поступила в редакцию 21.09.2020

METHOD OF STRAIGHTENING DISTORTED DUE TO MULTICOLLINEARITY COEFFICIENTS IN REGRESSION MODELS

M. P. Bazilevskiy

Department of Mathematics, Irkutsk State Transport University, 15 Chernyshevskogo Str., Irkutsk 664074, Russian Federation

Abstract: When constructing regression models, due to the strong multicollinearity of the explanatory variables, its coefficients are distorted, in particular, their signs, which negatively affects the interpretational qualities of such regression. This article is devoted to the development of a method of straightening coefficients distorted due to multicollinearity. This method is based on the property of the fully connected linear regression models proposed by the author. A nonlinear system, which is used to estimate fully connected regressions, is investigated. It is shown that the solution of this system can be obtained numerically using the method of simple iterations. A method for choosing unknown lambda-parameters in fully connected regression is proposed. It was found that in multivariate fully connected models with a strong correlation of all factors, the signs of the coefficients for the variables in the secondary equation coincide with the corresponding signs of the correlation coefficients. To straighten the distorted coefficients on the basis of this research, the "Selection B" algorithm was developed. The developed method of straightening has been successfully demonstrated by the example of modeling Russia's gross domestic product (GDP).

Keywords: regression analysis; fully connected linear regression model; multicollinearity; interpretation; numerical method; GDP of Russia

DOI: 10.14357/19922264210209

References

1. Gunst, R. F., and J. T. Webster. 1975. Regression analysis and problems of multicollinearity. *Commun. Stat. Theory* 4:277–292.
2. Tamura, R., K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, and T. Matsui. 2017. Best subset selection for eliminating multicollinearity. *J. Oper. Res. Soc. Jpn.* 60(3):321–336.
3. Förster, E., and B. Rönz. 1983. *Methoden der korrelation und regressionsanalyse*. Berlin: Verlag Die Wirtschaft, 1979. 369 p.
4. Chatterjee, S., and A. S. Hadi. 2012. *Regression analysis by example*. 5th ed. Hoboken, NJ: Wiley. 424 p.
5. Jolliffe, I. T. 2002. *Principal component analysis*. New York, NY: Springer-Verlag. 488 p.
6. Hoerl, A. E., and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55–67.
7. Mokshina, S. I., G. V. Shurshikova, and S. S. Shchekunskikh. 2017. Metod postroeniya soderzhatel'no interpretiruemykh regressionnykh modeley v usloviyakh mul'tikollinearnosti [The construction method of meaningful interpreted regression models in conditions of multicollinearity]. *Sovremennaya ekonomika: problemy i resheniya* [Modern Economics: Problems and Solutions] 89(5):81–94.
8. Bazilevskiy, M. P. 2019. Sintez modeli parnoy lineynoy regressii i prosteyshey EIV-modeli [Synthesis of linear regression model and EIV-model]. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii* [Modeling, Optimization and Information Technology] 7(1):170–182.
9. Bazilevskiy, M. P. 2019. Issledovanie dvukhfaktornoy modeli polnosvyaznoy lineynoy regressii [Investigation of a two-factor fully connected linear regression model]. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii* [Modeling, Optimization and Information Technology] 7(2):80–96.
10. Bazilevskiy, M. P. 2020. Mnogofaktornye modeli polnosvyaznoy lineynoy regressii bez ogranicheniy na sootnosheniya dispersiy oshibok peremennykh [Multifactor fully connected linear regression models without constraints to the ratios of variables errors variances]. *Informatika i ee Primeneniya — Inform. Appl.* 14(2):92–97.

Received September 21, 2020

Contributor

Bazilevskiy Mikhail P. (b. 1987) — Candidate of Science (PhD) in technology, associate professor, Irkutsk State Transport University, 15 Chernyshevkogo Str., Irkutsk 664074, Russian Federation; mik2178@yandex.ru

СОГЛАСОВАНИЕ ЦЕЛЕЙ АГЕНТОВ СПЛОЧЕННЫХ ГИБРИДНЫХ ИНТЕЛЛЕКТУАЛЬНЫХ МНОГОАГЕНТНЫХ СИСТЕМ*

И. А. Кириков[1], С. В. Листопад[2]

Аннотация: При разработке интеллектуальной системы как сообщества разнородных интеллектуальных агентов важна организация их взаимодействия. Для снижения трудоемкости этой процедуры предлагается методами сплоченных гибридных интеллектуальных многоагентных систем (СГИМАС) моделировать механизмы возникновения сплоченности в коллективах специалистов, решающих проблемы «за круглым столом». Агенты таких систем должны быть способны самостоятельно согласовывать цели, модели предметной области и вырабатывать протокол для решения поставленной проблемы. В статье предлагается модель согласования целей агентов СГИМАС.

Ключевые слова: сплоченность; гибридная интеллектуальная многоагентная система; коллектив специалистов

DOI: 10.14357/19922264210210

1 Введение

Решение проблем малыми коллективами специалистов снижает влияние человеческого фактора на вероятность ошибок, обеспечивая возможность комплексного, всестороннего рассмотрения проблемы, учета целей различных заинтересованных сторон. При этом для обеспечения эффективного коллективного решения проблем недостаточно собрать специалистов и обозначить проблему: полярность точек зрений, разнородность знаний, отсутствие принятых норм взаимодействия обусловливают напряженность и конфликтные ситуации. Для эффективной совместной деятельности малая группа в процессе своего развития должна пройти сложный путь от конгломерата незнакомых друг с другом специалистов без общей цели до сплоченного коллектива единомышленников, осуществляющих совместную деятельность и добивающихся результата на основе гармонизации целей, интересов и ценностей [1].

При моделировании гибридными интеллектуальными многоагентными системами (ГиИМАС) совместной работы специалистов по решению тех или иных проблем возникают аналогичные сложности. Агенты системы, собранные из разных репозиториев и построенные разными разработчиками, могут быть несовместимы по языкам передачи сообщений, целям, моделям предметной области или протоколам решения проблем. Для преодоления данных трудностей в [2] предложена модель СГИМАС. В их основу положена концепция функциональных гибридных интеллектуальных систем А. В. Колесникова [3], обеспечивающая учет проблемной неоднородности, многоагентный подход к построению интеллектуальных систем [4–7], позволяющий имитировать взаимодействие специалистов в коллективе, а также стратометрическая концепция (СК) сплоченности А. В. Петровского [8], описывающая условия формирования сплоченного коллектива агентов, понимающих друг друга, разделяющих общие цели и нормы. Цель настоящей работы — разработка модели согласования целей агентов СГИМАС.

2 Модель сплоченной гибридной интеллектуальной многоагентной системы

Предложенная в [2] СГИМАС моделирует сплоченность коллектива специалистов на двух из трех уровней СК А. В. Петровского [8] (из-за отсутствия эмоциональной составляющей у агентов уровень эмоциональных межличностных отношений СК не рассматривается):

(1) ценностно-ориентационное единство (ЦОЕ), т. е. близость основных ценностей и убежде-

*Исследование выполнено при финансовой поддержке РФФИ (проект 20-07-00104а).

[1]Калининградский филиал Федерального исследовательского центра «Информатика и управление» Российской академии наук, baltbipiran@mail.ru

[2]Калининградский филиал Федерального исследовательского центра «Информатика и управление» Российской академии наук, ser-list-post@yandex.ru

ний, возникающая в результате совместной деятельности;

(2) ядро — единство, обусловленное сходством целей членов коллектива.

Формально СГИМАС [2] описывается следующим выражением:

$$\text{chimas} = \\ = \langle \text{AG}, \text{env}, \text{INT}, \text{ORG}, \{\text{glng}, \text{ontng}, \text{protng}\}\rangle , \quad (1)$$

где AG — множество агентов системы, описываемых выражением (2), включающее подмножество $\text{AG}^{\text{sp}} \subseteq \text{AG}$ агентов-специалистов, моделирующих знания и рассуждения специалистов — членов коллектива, агента $\text{ag}^{\text{dm}} \in \text{AG}$, принимающего решения (АПР), агента-фасилитатора (АФ) $\text{ag}^{\text{fc}} \in \text{AG}$, отвечающего за организацию эффективного взаимодействия агентов системы и формирование сплоченности, а также служебных агентов, обеспечивающих взаимодействие агентов между собой; env — концептуальная модель внешней среды СГИМАС; INT — множество элементов для структурирования взаимодействий агентов (3); ORG — множество архитектур СГИМАС; $\{\text{glng}, \text{ontng}, \text{protng}\}$ — множество моделей макроуровневых процессов, содержащее модель glng согласования целей агентов, обеспечивающую сплоченность на уровне ядра СК, описываемую выражением (5), модель ontng согласования онтологий (моделей предметной области) агентов, соответствующая обмену знаниями, опытом и убеждениями между членами коллектива и формированию сплоченности на уровне ЦОЕ, и модель protng построения агентами протокола сплоченного решения проблем, имитирующая выработку и интериоризацию членами коллектива норм взаимодействия на уровне ЦОЕ.

Агент $\text{ag} \in \text{AG}$ из формулы (1) описывается выражением:

$$\text{ag} = \langle \text{id}^{\text{ag}}, \text{gl}^{\text{ag}}, \text{LANG}^{\text{ag}}, \text{ont}^{\text{ag}}, \text{ACT}^{\text{ag}}, \text{prot}^{\text{ag}}\rangle, \quad (2)$$

где id^{ag} — идентификатор (имя) агента; gl^{ag} — цель агента в виде нечеткого множества с функцией принадлежности $\mu_{\text{id}}(\text{pr}^{\text{cs}}_{\text{id}\,1}, \ldots \text{pr}^{\text{cs}}_{\text{id}\,\text{Nprid}})$, заданной на подмножестве концептов-свойств $\text{PR}^{\text{cs}}_{\text{id}} = \{\text{pr}^{\text{cs}}_{\text{id}\,1}, \ldots, \text{pr}^{\text{cs}}_{\text{id}\,\text{Nprid}}\}$ множества концептов $\text{PR}^{\text{cs}}_{\text{id}} \subseteq \text{PR}_{\text{id}} \subseteq C_{\text{id}}$ онтологии агента ont^{ag}; $\text{LANG}^{\text{ag}} \subseteq \text{LANG}$ — множество языков передачи сообщений, которыми «владеет» агент; ont^{ag} — модель предметной области (онтология) агента, описываемая выражением (4); ACT^{ag} — множество действий, реализуемых агентом; prot^{ag} — модель протокола решения проблемы, разрабатываемая агентом в процессе взаимодействия с другими агентами [4].

Множество элементов для структурирования взаимодействий агентов из формулы (1) описывается выражением:

$$\text{INT} = \left\{\text{prot}^{\text{bsc}}, \text{PRC}, \text{LANG}, \text{ont}^{\text{bsc}}, \text{chn}\right\} , \quad (3)$$

где prot^{bsc} — базовый протокол, обеспечивающий взаимодействие агентов по формированию протокола сплоченного взаимодействия для решения поставленных перед СГИМАС проблем; PRC — множество элементов для конструирования протокола решения проблем агентами-специалистами и АПР; LANG — множество языков передачи сообщений агентов; ont^{bsc} — базовая онтология, обеспечивающая интерпретацию агентами семантики сообщений по согласованию собственных целей и моделей предметной области, формированию протокола сплоченного взаимодействия, описываемая выражением (4); chn — степень сплоченности агентов [2], описывающая степень сходства целей и онтологий, а также согласованности протокола решения проблем.

Модели онтологий агентов ont^{ag} и базовой онтологий ont_{bsc} из выражений (2) и (3) соответственно описываются следующим образом:

$$\text{ont} = \\ = \langle L, C, R, \text{AT}, \text{FC}, \text{FR}, \text{FA}, H^c, H^r, \text{INST}\rangle , \quad (4)$$

где $L = L^c \cup L^r \cup L^{\text{at}} \cup L^{\text{va}}$ — лексикон, множество лексем, состоящее из подмножеств лексем, обозначающих понятия L^c, отношения L^r, атрибуты L^{at} и их значения L^{va}; C — множество концептов (понятий); $R : C \times C$ — множество отношений между концептами, первая компонента кортежа отношения называется доменом $\text{dm}\,(r) = \text{Пр}_1(r)$, а вторая — диапазоном значений отношения $\text{rn}\,(r) = \text{Пр}_2(r)$; $\text{AT} : C \times L^{\text{va}}$ — множество атрибутов концептов; $\text{FC} : 2^{L^c} \to 2^C$ — функция связи лексикона с концептами; $\text{FR} : 2^{L^r} \to 2^R$ — функция связи лексикона с отношениями; $\text{FA} : L^{\text{at}} \to \\ \to \text{AT}$ — функция связи лексикона с атрибутами; $H^c = C \times C$ — таксономическая иерархия концептов; $H^r = R \times R$ — иерархия отношений; INST — множество экземпляров, концептов единичного объема [9]. Функции FC и FR предполагают, что в общем случае одна лексема может соответствовать нескольким концептам или отношениям и, наоборот, один концепт или отношение может описываться несколькими лексемами.

3 Модель согласования целей агентов

Модель согласования целей агентов описывается выражением:

$$\mathrm{glng} = \langle \mathrm{glest}, \mathrm{glngn}, \mathrm{GLNM} \rangle , \qquad (5)$$

где glest — модель оценки сходства целей; glngn — модель оценки необходимости согласования целей; $\mathrm{GLNM} = \{\mathrm{glneg}, \mathrm{glarg}, \mathrm{gldmo}\}$ — множество методов согласования, например путем споров glarg, переговоров glneg или на основе распоряжений АПР gldmo.

Модель оценки сходства целей пары агентов ag_i и ag_j, которая подробно рассмотрена в [10], может быть представлена следующим выражением:

$$\begin{aligned}\mathrm{glest} = {} & r_1^{\mathrm{act\,act}}\left(\mathrm{act}_{\mathrm{cm}}^{\mathrm{ag}}, \mathrm{act}_{\mathrm{cu}}^{\mathrm{ag}}\right) \circ r_1^{\mathrm{act\,act}}\left(\mathrm{act}_{\mathrm{cu}}^{\mathrm{ag}}, \mathrm{act}_{\mathrm{cvr}}^{\mathrm{ag}}\right) \circ \\ & \circ r_1^{\mathrm{act\,act}}\left(\mathrm{act}_{\mathrm{cvr}}^{\mathrm{ag}}, \mathrm{act}_{\mathrm{gsmc}}^{\mathrm{ag}}\right) \circ r_1^{\mathrm{act\,pr}}\left(\mathrm{act}_{\mathrm{cm}}^{\mathrm{ag}}, \mathrm{PR}_i^{\mathrm{cs}}\right) \circ \\ & \circ r_1^{\mathrm{act}\,c}\left(\mathrm{act}_{\mathrm{cm}}^{\mathrm{ag}}, C_j\right) \circ r_2^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cm}}^{\mathrm{ag}}, \mathrm{MP}_{i\,j}\right) \circ \\ & \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cu}}^{\mathrm{ag}}, \mathrm{MP}_{i\,j}\right) \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cu}}^{\mathrm{ag}}, \mathrm{ont}_i^{\mathrm{ag}}\right) \circ \\ & \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cu}}^{\mathrm{ag}}, \mathrm{ont}_j^{\mathrm{ag}}\right) \circ r_2^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cu}}^{\mathrm{ag}}, \mathrm{MP}''_{i\,j}\right) \circ \\ & \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cvr}}^{\mathrm{ag}}, \mathrm{MP}''_{i\,j}\right) \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cvr}}^{\mathrm{ag}}, \mu_i\right) \circ \\ & \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cvr}}^{\mathrm{ag}}, \mu_j\right) \circ r_2^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cvr}}^{\mathrm{ag}}, \mu'_i\right) \circ \\ & \circ r_2^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{cvr}}^{\mathrm{ag}}, \mu'_j\right) \circ \\ & \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{gsmc}}^{\mathrm{ag}}, \mu'_i\right) \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{gsmc}}^{\mathrm{ag}}, \mu'_j\right) \circ \\ & \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{gsmc}}^{\mathrm{ag}}, \mathrm{MP}''_{i\,j}\right) \circ r_2^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{gsmc}}^{\mathrm{ag}}, \mathrm{gls}_{i\,j}^{\mathrm{ag}}\right) ,\end{aligned}$$

где $r_1^{\mathrm{act\,act}}$ — отношение «следование» типа «действие–действие» [3]; $\mathrm{act}_{\mathrm{cm}}^{\mathrm{ag}}$ — действие по установлению соответствия $\mathrm{MP}_{i\,j}$ (сходство пары концептов определяется как среднее геометрическое мер лексикографического и таксономического сходства [11]) между концептами-свойствами $\mathrm{PR}_i^{\mathrm{cs}}$, на которых определена нечеткая цель агента ag_i, и концептами C_j онтологии агента ag_j, поскольку в онтологиях агентов идентификаторы концептов-свойств, на которых определены цели, и их число могут различаться; $\mathrm{act}_{\mathrm{cu}}^{\mathrm{ag}}$ — действие по выявлению, объединению и сокращению функционально зависимых концептов в соответствии $\mathrm{MP}_{i\,j}$, в результате чего строится модифицированное соответствие $\mathrm{MP}''_{i\,j}$ независимых концептов обеих онтологий, на которых определены нечеткие цели агентов ag_i и ag_j; $\mathrm{act}_{\mathrm{cvr}}^{\mathrm{ag}}$ — действие по замене переменных в функциях принадлежности μ_i и μ_j нечетких целей агентов, в результате чего формируются модифицированные функции принадлежности μ'_i и μ'_j; $\mathrm{act}_{\mathrm{gsmc}}^{\mathrm{ag}}$ — действие по расчету значения меры сходства $\mathrm{gls}_{i\,j}^{\mathrm{ag}}$ нечетких целей [10] с учетом степени сходства концептов-свойств, на которых они определены; $r_1^{\mathrm{act\,pr}}$ — отношение «иметь аргументом» типа «действие–свойство»; $r_1^{\mathrm{act}\,c}$ — отношение «иметь аргументом» типа «действие–концепт»; $r_2^{\mathrm{act\,res}}$ — отношение «иметь результатом» типа «действие–ресурс»; $r_1^{\mathrm{act\,res}}$ — отношение «иметь аргументом» типа «действие–ресурс»; $\circ$ — операция склеивания концептов [3].

Необходимость согласования целей агентов оценивается АФ в соответствии со своей нечеткой базой знаний, представленной в [12]. Она позволяет АФ организовать работу агентов системы в соответствии с моделью ромба группового принятия решений С. Кейнера [13], содержащей три последовательные фазы: дивергентное коллективное мышление, в ходе которого вырабатываются альтернативные решения проблемы, стадию бурления, на которой необходимо повышать «взаимопонимание» между агентами, сближать их цели, модели предметной области и вырабатывать согласованный протокол решения поставленной проблемы, и стадию конвергентного мышления, когда предложенные альтернативы классифицируются, ранжируются и дорабатываются для принятия интегрированного, устраивающего всех агентов решения. Модель оценки необходимости согласования целей агентов описывается выражением:

$$\begin{aligned}\mathrm{glngn} = {} & r_1^{\mathrm{act\,act}}\left(\mathrm{act}_{\mathrm{dmsa}}^{\mathrm{ag}}, \mathrm{act}_{\mathrm{gnmc}}^{\mathrm{ag}}\right) \circ \\ & \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{dmsa}}^{\mathrm{ag}}, \mathrm{MSG}^{\mathrm{sol}}\right) \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{dmsa}}^{\mathrm{ag}}, \mathrm{GL}^{\mathrm{ag}}\right) \circ \\ & \circ r_1^{\mathrm{act\,st}}\left(\mathrm{act}_{\mathrm{dmsa}}^{\mathrm{ag}}, \mathrm{pss}^{\mathrm{it}}\right) \circ r_2^{\mathrm{act\,st}}\left(\mathrm{act}_{\mathrm{dmsa}}^{\mathrm{ag}}, \mathrm{pss}^{\mathrm{it}+1}\right) \circ \\ & \circ r_1^{\mathrm{act\,st}}\left(\mathrm{act}_{\mathrm{gnmc}}^{\mathrm{ag}}, \mathrm{pss}^{\mathrm{it}+1}\right) \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{gnmc}}^{\mathrm{ag}}, \mathrm{GLNM}\right) \circ \\ & \circ r_1^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{gnmc}}^{\mathrm{ag}}, \mathrm{glnm}^{\mathrm{it}}\right) \circ r_2^{\mathrm{act\,res}}\left(\mathrm{act}_{\mathrm{gnmc}}^{\mathrm{ag}}, \mathrm{glnm}^{\mathrm{it}+1}\right),\end{aligned}$$

где $\mathrm{act}_{\mathrm{dmsa}}^{\mathrm{ag}}$ — действие по вычислению нечеткой переменной «состояние процесса коллективного решения проблемы» $\mathrm{pss}^{\mathrm{it}+1}$ на основе множества $\mathrm{MSG}^{\mathrm{sol}} \subseteq \mathrm{MSG} \subseteq \mathrm{INST}$ сообщений — решений проблемы или ее частей, степени близости целей агентов $\mathrm{GL}^{\mathrm{ag}}$ и значения переменной на предыдущей итерации $\mathrm{pss}^{\mathrm{it}}$; $\mathrm{act}_{\mathrm{gnmc}}^{\mathrm{ag}}$ — действие по выбору метода $\mathrm{glnm}^{\mathrm{it}+1}$ согласования целей на новой итерации процесса решения проблемы из множества GLNM на основании метода на текущей итерации $\mathrm{glnm}^{\mathrm{it}}$ и вычисленного значения нечеткой переменной «состояние процесса коллективного решения проблемы» $\mathrm{pss}^{\mathrm{it}+1}$; $r_1^{\mathrm{act\,st}}$ — отношение «иметь аргументом» типа «действие–состояние»; $r_2^{\mathrm{act\,st}}$ — отношение «иметь результатом» типа «действие–состояние».

Методы согласования целей из множества GLNM описывают механизмы изменения целей агентов в процессе их «общения» в форме обмена сообщениями путем споров, переговоров или на основе распоряжений АПР. Споры и переговоры необходимы для повышения сплоченности, когда знания агентов друг о друге или о решаемой проблеме неполны [14]. Корректировка на основе распоряжений АПР имеет серьезный недостаток: АПР вмешивается в систему целеполагания

агентов, моделирующих знания реальных специалистов по решаемой проблеме, что может привести к нерелевантности предлагаемых СГИМАС решений точкам зрения на проблему моделируемых специалистов. В связи с этим данный метод может быть задействован, только если АПР может получить от агентов достоверные сведения об их целях, при этом споры и переговоры не привели к желаемому результату, т. е. после проведения споров и переговоров АФ оценивает «состояние процесса коллективного решения проблемы» $\mathrm{pss}^{\mathrm{it}+1}$ как требующее согласования целей агентов. Необходимое условие для выполнения любого из методов согласования целей агентов — согласование онтологий агентов в пределах верхних котопий концептов (множества концептов, содержащего все вышележащие концепты по таксономической иерархии концептов H^c по отношению к заданному концепту и сам концепт [11]), на которых определены нечеткие цели.

В ходе споров glarg агенты обмениваются сообщениями-аргументами, которые направлены на изменение целевой функции агента-адресата. На основе анализа рассмотренных в [14] типов аргументов, применяемых в переговорах специалистов, для согласования целей агентами СГИМАС предлагается использовать следующие механизмы аргументации:

- примеры и контрпримеры, демонстрирующие противоречие между целями агента-адресата и результатами реализации предыдущих его предложений;
- обращение к «сложившейся практике», демонстрирующее, что агенты, ранее выполнявшие в СГИМАС роль, занимаемую агентом-адресатом, придерживались предлагаемой цели, что способствовало высокому качеству коллективных решений;
- апелляция к коллективным целям, чтобы убедить агента-адресата, что корректировка его цели позволит принять решение, соответствующее поставленным перед системой целям.

Получив аргументированное предложение по корректировке своей цели, агент-адресат оценивает изложенные аргументы с использованием функции «анализ аргументов» и в случае согласия с ними корректирует свою цель в соответствии с полученным предложением.

Переговоры агентов по поводу согласования целей glneg заключаются в формировании сообщений-запросов на корректировку целей и угроз по корректировке собственной цели в сторону, невыгодную адресату сообщения, если предложение будет отвергнуто, или уступок, выгодных для адресата, если предложение будет принято. Для этого может использоваться метод монотонных минимальных уступок [15]: агенты поочередно отправляют сообщения-предложения, начиная с самых выгодных для себя, и в процессе переговоров монотонно отступают от своих первоначальных требований. В результате множество возможных соглашений относительно целей агентов оказывается состоящим из всех индивидуально рациональных соглашений, эффективных по Парето [4].

В случае необходимости корректировки целей на основе распоряжений АПР агенты-специалисты отправляют ему сообщения о своих целях на текущий момент. Агент, принимающий решения, анализирует множество целей агентов, сопоставляя со своей, и формирует распоряжения для каждого из агентов по корректировке его функции принадлежности нечеткой цели. Получив такое распоряжение, агенты-специалисты корректируют свои цели без дальнейших обсуждений.

Таким образом, предложенная модель и методы согласования целей агентов СГИМАС позволят снизить интенсивность конфликтов, обусловленных различиями в целях агентов, созданных разными группами разработчиков и моделирующих различных специалистов по решаемой проблеме. Благодаря наличию модели оценки необходимости согласования целей АФ может инициировать релевантные ситуации механизмы согласования целей и приостанавливать их использование для предотвращения таких нежелательных эффектов от чрезмерной сплоченности, как конформизм.

4 Заключение

Рассмотрены особенности распределенной разработки систем гибридного и синергетического искусственного интеллекта на примере ГиИМАС. С целью снижения трудозатрат на интеграцию автономных частей интеллектуальной системы предложено моделирование механизмов сплочения коллектива, возникающих в длительно существующих группах специалистов, решающих практические проблемы «за круглым столом». Для этих целей предложено разработать новый класс интеллектуальных систем — сплоченные гибридные интеллектуальные системы. Рассмотрена модель системы такого класса и разработана модель согласования целей ее агентов. Механизм согласования целей в ходе решения проблемы позволит агенту вырабатывать решения с учетом не только собственных целей, заложенных разработчиками при моделировании знаний и поведения соответствующего специалиста, но и, хотя бы частично, учитывать цели

агентов, моделирующих других специалистов. Это позволит снизить вероятность досрочного завершения коллективного решения проблемы и принятия никого не устраивающего решения из-за несовместимости целей и точек зрения на проблему.

Литература

1. *Почебут Л. Г., Чикер В. А.* Организационная социальная психология. — СПб.: Речь, 2002. 298 с.
2. *Listopad S.* Modeling team cohesion using hybrid intelligent multi-agent systems // 2nd Conference (International) on Control Systems, Mathematical Modeling, Automation and Energy Efficiency Proceedings. — Piscataway, NJ, USA: IEEE, 2020. P. 416–421.
3. *Колесников А. В., Кириков И. А., Листопад С. В.* Гибридные интеллектуальные системы с самоорганизацией: координация, согласованность, спор. — М.: ИПИ РАН, 2014. 189 с.
4. *Тарасов В. Б.* Агенты, многоагентные системы, виртуальные сообщества: стратегическое направление в информатике и искусственном интеллекте // Новости искусственного интеллекта, 1998. № 2. С. 5–63.
5. *Городецкий В. И., Грушинский М. С., Хабалов А. В.* Многоагентные системы (обзор) // Новости искусственного интеллекта, 1998. № 2. С. 64–116.
6. *Хорошевский В. Ф.* Поведение интеллектуальных агентов: модели и методы реализации // 4-й Междунар. семинар по прикладной семиотике, семиотическому и интеллектуальному управлению: Сб. научных трудов. — Переславль-Залесский: РАИИ, 1999. С. 5–20.
7. *Wooldridge M.* An introduction to multiagent systems. — New York, NY, USA: Wiley, 2009. 484 p.
8. *Петровский А. В.* Опыт построения социально-психологической концепции групповой активности // Вопросы психологии, 1973. № 5. С. 3–17.
9. *Крюков К. В., Панкова Л. А., Пронина В. А., Суховеров В. С., Шипилина Л. Б.* Меры семантической близости в онтологии // Проблемы управления, 2010. № 5. С. 2–14.
10. *Listopad S.* Estimating of the similarity of agents' goals in cohesive hybrid intelligent multi-agent system // CEUR Workshop Proceedings, 2020. Vol. 2782. P. 180–185.
11. *Maedche A., Zacharias V.* Clustering ontology-based metadata in the semantic web // Principles of data mining and knowledge discovery / Eds. T. Elomaa, H. Mannila, H. Toivonen. — Lecture notes in artificial intelligence. — Springer, 2002. Vol. 2431. P. 348–360.
12. *Листопад С. В., Румовская С. Б.* Нечеткое управление гетерогенным мышлением агентов гибридной интеллектуальной многоагентной системы // Системы и средства информатики, 2020. Т. 30. № 4. С. 38–49.
13. *Kaner S., Lind L., Toldi C., Fisk S., Beger D.* The facilitator's guide to participatory decision-making. — San Francisco, CA, USA: Jossey-Bass, 2011. 368 p.
14. *Kraus S., Sycara K., Evenchik A.* Reaching agreements through argumentation: A logical model and implementation // Artif. Intell., 1998. Vol. 104. P. 1–60.
15. *Rosenshein J., Zlotkin G.* Rules of encounter: Designing conventions for automated negociation among computers. — Cambridge, MA, USA: MIT Press, 1994. 253 p.

Поступила в редакцию 05.04.2021

COORDINATION OF AGENTS' GOALS IN COHESIVE HYBRID INTELLIGENT MULTIAGENT SYSTEMS

I. A. Kirikov and S. V. Listopad

Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation

Abstract: When developing an intelligent system as a community of heterogeneous intelligent agents, it is important to organize their interaction. To reduce the complexity of this procedure, it is proposed to simulate with methods of cohesive hybrid intelligent multiagent systems the mechanisms of cohesion emergence in teams of specialists solving problems "at a round table." Agents of such systems should be able to independently coordinate their goals and domain models and develop a protocol to solve the posed problem. The article proposes a model for coordinating the goals of agents of cohesive hybrid intelligent multiagent systems.

Keywords: cohesion; hybrid intelligent multiagent system; team of specialists

DOI: 10.14357/19922264210210

Acknowledgments

The reported study was funded by RFBR, project number 20-07-00104a.

References

1. Pochebut, L. G., and V. A. Chiker. 2002. *Organizatsionnaya sotsial'naya psikhologiya* [Organizational social psychology]. St. Petersburg: Rech. 298 p.
2. Listopad, S. 2020. Modeling team cohesion using hybrid intelligent multi-agent systems. *2nd Conference (International) on Control Systems, Mathematical Modeling, Automation and Energy Efficiency Proceedings*. Piscataway, NJ: IEEE. 416–421.
3. Kolesnikov, A. V., I. A. Kirikov, and S. V. Listopad. 2014. *Gibridnye intellektual'nye sistemy s samoorganizatsiey: koordinatsiya, soglasovannost', spor* [Hybrid intelligent systems with self-organization: Coordination, consistency, dispute]. Moscow: IPI RAN. 189 p.
4. Tarasov, V. B. 1998. Agenty, mnogoagentnye sistemy, virtual'nye soobshchestva: strategicheskoe napravlenie v informatike i iskusstvennom intellekte [The agents, multiagent system, virtual communities: Strategic direction in computer science and artificial intelligence]. *Novosti iskusstvennogo intellekta* [News of Artificial Intelligence] 2:5–63.
5. Gorodetskiy, V. I., M. S. Grushinskiy, and A. V. Khabalov. 1998. Mnogoagentnye sistemy (obzor) [Multi-agent systems (review)]. *Novosti iskusstvennogo intellekta* [News of artificial intelligence] 2:64–116.
6. Khoroshevskiy, V. F. 1999. Povedenie intellektual'nykh agentov: modeli i metody realizatsii [The behavior of intelligent agents: Models and methods of implementation]. *4th Workshop (International) on Applied Semiotics, Semiotics and Intelligent Management Proceedings*. Pereslavl'-Zalesskiy: RAAI. 5–20.
7. Wooldridge, M. 2009. *An introduction to multiagent systems*. New York, NY: Wiley. 484 p.
8. Petrovskiy, A. V. 1973. Opyt postroeniya sotsial'no-psikhologicheskoy kontseptsii gruppovoy aktivnosti [The experience of building a socio-psychological concept of group activity]. *Voprosy psikhologii* [Psychology Issues] 5:3–17.
9. Kryukov, K. V., L. A. Pankova, V. A. Pronina, V. S. Sukhoverov, and L. B. Shipilina. 2010. Mery semanticheskoy blizosti v ontologii [Measures of semantic proximity in ontology]. *Problemy upravleniya* [Control Sciences] 5:2–14.
10. Listopad, S. 2020. Estimating of the similarity of agents' goals in cohesive hybrid intelligent multi-agent system. *CEUR Workshop Proceedings*. 2782:180–185.
11. Maedche, A., and V. Zacharias. 2002. Clustering ontology-based metadata in the semantic web. *Principles of data mining and knowledge discovery*. Eds. T. Elomaa, H. Mannila, and H. Toivonen. Lecture notes in artificial intelligence ser. Springer. 2431:348–360.
12. Listopad, S. V., and S. B. Rumovskaya. 2020. Nechetkoe upravlenie geterogennym myshleniem agentov gibridnoy intellektual'noy mnogoagentnoy sistemy [Fuzzy control of heterogeneous thinking of the hybrid intelligent multiagent system's agents]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 30(4):38–49.
13. Kaner, S., L. Lind, C. Toldi, S. Fisk, and D. Beger. 2011. *The facilitator's guide to participatory decision-making*. San Francisco, CA: Jossey-Bass. 368 p.
14. Kraus, S., K. Sycara, and A. Evenchik. 1998. Reaching agreements through argumentation: A logical model and implementation. *Artif. Intell.* 104:1–60.
15. Rosenshein, J., and G. Zlotkin. 1994. *Rules of encounter: Designing conventions for automated negotiation among computers*. Cambridge, MA: MIT Press. 253 p.

Received April 5, 2021

Contributors

Kirikov Igor A. (b. 1955) — Candidate of Science (PhD) in technology, director, Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation; baltbipiran@mail.ru

Listopad Sergey V. (b. 1984) — Candidate of Science (PhD) in technology, senior scientist, Kaliningrad Branch of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 5 Gostinaya Str., Kaliningrad 236000, Russian Federation; ser-list-post@yandex.ru

РАЗЛОЖЕНИЯ ЧЕБЫШЁВА–ЭДЖВОРТА ДЛЯ РАСПРЕДЕЛЕНИЙ ОБОБЩЕННЫХ СТАТИСТИК ТИПА ХОТЕЛЛИНГА, ПОСТРОЕННЫХ ПО ВЫБОРКАМ СЛУЧАЙНОГО РАЗМЕРА*

М. М. Монахов[1]

Аннотация: Доказан аналог теоремы переноса для функций распределения статистики типа Хотеллинга, размер которой является случайной величиной, позволяющий оценить скорость сходимости разложения Чебышёва–Эджворта и получить явный вид вышеупомянутого разложения для исходной статистики. На основании следствия к доказанному аналогу теоремы переноса для случая, когда размер статистики имеет отрицательное биномиальное распределение (смещенное на 1), получен явный вид разложения Чебышёва–Эджворта второго порядка на базе предельного F-распределения. По построенному разложению Чебышёва–Эджворта для специального значения параметра случайного размера выборки построено разложение Корниша–Фишера второго порядка на базе квантилей F-распределения. Проведен вычислительный эксперимент и построены графики, иллюстрирующие полученное разложение Чебышёва–Эджворта.

Ключевые слова: обобщенные разложения Чебышёва–Эджворта; разложения Корниша–Фишера; выборка случайного объема; F-распределение; статистика типа Хотеллинга

DOI: 10.14357/19922264210211

1 Введение

В анализе данных довольно часто возникает задача множественных сравнений. Например, различных возрастных, профессиональных, социальных слоев населения, или влияния различных доз препарата, методов диагностики и т. д. Данную задачу помогает решить дисперсионный анализ, который применяется для исследования влияния одной или нескольких качественных переменных (факторов) на одну зависимую количественную переменную. Дисперсионный анализ широко применяется в сфере производства, здравоохранения, рекламы, продовольствия, обслуживания, его реализации представлены в статистических пакетах для многих языков программирования. Сущность дисперсионного анализа заключается в расчленении общей дисперсии изучаемого признака на отдельные компоненты, обусловленные влиянием конкретных факторов, и проверке гипотез о значимости влияния этих факторов на исследуемый признак. Дополнительные проблемы возникают в случае, когда объем наблюдения оказывается случайным [1].

В задачах многомерного однофакторного дисперсионного анализа рассматриваются q выборок с фиксированным размером $n_1, \ldots, n_q$: $(X_{11}, \ldots, X_{1n_1}), \ldots, (X_{q1}, \ldots, X_{qn_q})$, где X_{ij} — p-мерное наблюдение, представимое в виде:

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}\,.$$

Здесь μ и α_i — неизвестные векторные параметры; ϵ_{ij} — случайные ошибки, являющиеся независимыми одинаково распределенными случайными величинами с нормальным распределением $N_p(0, B)$. При рассмотрении основной гипотезы однородности выборок

$$H_0 : \alpha_1 = \cdots = \alpha_q = 0$$

определяются матрицы S_h и S_e, отражающие межуровневые и внутриуровневые различия соответственно для элементов выборок

$$S_h = \sum_{i=1}^{q} n_i(\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})';$$

$$S_e = \sum_{i=1}^{q}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'$$

с $n = n_1 + \cdots + n_q$ и

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{q}\sum_{j=1}^{n_i} y_{ij}.$$

*Исследование выполнено в соответствии с программой Московского центра фундаментальной и прикладной математики.

[1]Московский центр фундаментальной и прикладной математики Московского государственного университета имени М. В. Ломоносова, mih_monah@mail.ru.

В предположении справедливости основной гипотезы H_0 случайные матрицы S_h и S_e независимы и имеют центральные распределения Уишарта $W_p(q, I_p)$ и $W_p(n, I_p)$ соответственно. На базе матриц S_h и S_e для проверки гипотезы H_0 строятся статистики, одной из которых является статистика Лоули–Хотеллинга.

В работах [1, 2] была доказана общая теорема переноса, позволяющая оценить скорость сходимости разложения типа Чебышёва–Эджворта первого порядка для асимптотически нормальных статистик, построенных по выборкам случайного объема, а также получить явный вид данного разложения. В качестве примера статистики в этих работах рассматривается выборочное среднее, которое приближается нормальным распределением. В работе [3] получено разложение Корниша–Фишера первого порядка для разложения Чебышёва–Эджворта из работы [1]. В работе [4] получены разложения Чебышёва–Эджворта и Корниша–Фишера второго порядка для статистик типа выборочного среднего, построенных по выборкам случайного объема. Данная работа развивает результаты вышеперечисленных работ. Для статистики типа Хотеллинга случайного размера доказан аналог теоремы переноса и построено асимптотическое разложение типа Чебышёва–Эджворта для функции распределения данной статистики.

Используем следующие обозначения: $\mathbf{R}$ — множество вещественных чисел; $\mathbf{N} := \{1, 2, \ldots\}$ — положительные целые числа; $\mathbf{I}_A(x)$ — индикаторная функция.

Определим статистику Лоули–Хотеллинга (см., например, [5]):

$$T_n = T_0^2 = n \operatorname{tr} S_h S_e^{-1}. \tag{1}$$

Рассмотрим случай, когда параметр n не определен заранее, а является случайной величиной N_n. В этом случае размеры выборок $n_1, \ldots, n_q$ становятся независимыми одинаково распределенными случайными величинами $N_{n_1}, \ldots, N_{n_q}$, а случайная матрица S_e становится случайной матрицей S_{N_n}, которая в предположении справедливости основной гипотезы H_0 имеет центральное распределение Уишарта $W_p(N_n, I_p)$. Обобщенная нормированная статистика Хотеллинга случайного размера запишется в виде:

$$T_{N_n} = \widetilde{T}_0^2 = g(n) \operatorname{tr} S_h S_{N_n}^{-1}. \tag{2}$$

В разд. 2 получен аналог теоремы переноса для обобщенной нормированной статистики Хотеллинга случайного размера, в разд. 3 построен аналог разложения Чебышёва–Эджворта для данной статистики, в разд. 4 получен явный вид разложения Корниша–Фишера для частного случая параметра размера данной статистики. В разд. 5 приведены доказательства полученных теорем.

2 Аналог теоремы переноса для статистики типа Хотеллинга

Запишем следующую теорему из работы [6, теорема 4.1].

Теорема 1. *Пусть статистика T_n определена в формуле* (1)*, $G_k(x) = \Pr\{\chi^2 < x\}$ — функция распределения хи-квадрат с k степенями свободы. Существует вещественное число $C_1 > 0$ такое, что для всех целых $n \geq 1$*

$$\sup_x \left| \mathbb{P}\left(n \operatorname{tr} S_h S_e^{-1} \leq x\right) - G_k(x) - \right.$$
$$\left. - \frac{k}{4n} \sum_{j=0}^{2} a_j G_{k+2j}(x) \right| \leq C_1 n^{-2}, \tag{3}$$

где $k = pq$; $a_0 = q - p - 1$; $a_1 = -2q$; $a_2 = q + p + 1$.

Предположим, что функция распределения нормированного случайного размера выборки N_n удовлетворяет следующему условию.

Условие 1. Существуют константы $m \in \mathbb{N}$, $\beta > m/2$, $C_2 > 0$, функция распределения $H(y)$ с $H(0+) = 0$, функции ограниченной вариации $h_i(y)$, $i = 1, \ldots, m$, последовательность $0 < g(n) \uparrow \infty$, $n \to \infty$ такие, что для всех целых $n \geq 1$

$$\sup_{y \geq 0} \left| \mathbb{P}\left(\frac{N_n}{g(n)} \leq y\right) - H(y) - \sum_{i=1}^{m} \frac{1}{n^{i/2}} h_i(y) \right| \leq$$
$$\leq C_2 n^{-\beta}, \quad n \in \mathbb{N}. \tag{4}$$

Сформулируем аналог теоремы переноса, позволяющий оценить распределение обобщенной нормированной статистики Хотеллинга случайного размера $g(n) \operatorname{tr} S_h S_{N_n}^{-1}$.

Теорема 2. *Пусть статистика T_{N_n} определена в формуле* (2) *и для случайного размера выборки N_n выполнено условие* 1*. Тогда существует константа $C_3 > 0$ такая, что справедливо неравенство*

$$\sup_x \left| \mathbb{P}\left(g(n) \operatorname{tr} S_h S_{N_n}^{-1} \leq x\right) - F_n(x) \right| \leq$$
$$\leq C_1 \mathbb{E} N_n^{-2} + \frac{C_3 + C_2 M_n}{n^{\beta}},$$

где

$$F_n(x) = \int\limits_{1/g(n)}^{\infty} G_k(xy)\,dH(y) + \\ + \sum_{i=1}^{m} \frac{1}{n^{i/2}} \int\limits_{1/g(n)}^{\infty} G_k(xy)\,dh_i(y) + \\ + \frac{k}{4g(n)} \int\limits_{1/g(n)}^{\infty} \sum_{j=0}^{2} \frac{a_j}{y}\, G_{k+2j}(xy)\,dH(y) + \\ + \frac{k}{4g(n)} \sum_{i=1}^{m} \frac{1}{n^{i/2}} \int\limits_{1/g(n)}^{\infty} \frac{1}{y} \sum_{j=0}^{2} a_j G_{k+2j}(xy)\,dh_i(y);$$

$$M_n = \sup_x \int\limits_{1/g(n)}^{\infty} \left| \frac{\partial}{\partial y} \left(G_k\,(yx) + \right.\right. \\ \left.\left. + \frac{k}{4g(n)y} \sum_{j=0}^{2} a_j G_{k+2j}\,(yx) \right) \right| dy\,.$$

Следствие 1. В условиях теоремы 2 с дополнительными предположениями

$$h_2(0) = 0\,;\;\; H\left(\frac{1}{g(n)}\right) \le c_0 n^{-\gamma}\,; \\ h_2\left(\frac{1}{g(n)}\right) \le c_1 n^{1-\gamma}; \\ \int\limits_0^{1/g(n)} \frac{1}{y}\,dH(y) \le c_2 g(n) n^{-\gamma}, \\ \int\limits_0^{1/g(n)} \frac{1}{y}\,dh_2(y) \le c_3 g(n) n^{1-\gamma}$$

для некоторого $\gamma > 1$ существует $C_3 = C_3(C_2, k, q)$ такая, что $\forall\, n \in \mathbb{N}$

$$\sup_x \left| \mathbb{P}\left(g(n)\,\mathrm{tr}\, S_h S_{N_n}^{-1} \le x\right) - F_{2;n}(x) \right| \le \\ \le C_1 \mathbb{E} N_n^{-2} + C_3 n^{-\min(\beta,\gamma)},$$

где

$$F_{2;n}(x) = \int\limits_0^{\infty} G_k(xy)\,dH(y) + \\ + \frac{k}{4g(n)} \int\limits_0^{\infty} \sum_{j=0}^{2} \frac{a_j}{y} G_{k+2j}(xy)\,dH(y) + \\ + \frac{1}{n} \int\limits_0^{\infty} G_k(xy)\,dh_2(y) + \\ + \frac{k}{4ng(n)} \int\limits_0^{\infty} \sum_{j=0}^{2} \frac{a_j}{y}\, G_{k+2j}(xy)\,dh_2(y)\,. \quad (5)$$

3 Разложение Чебышёва–Эджворта

Рассмотрим теперь пример применения теоремы 2. Пусть размер выборки $N_n(r)$ имеет отрицательное биномиальное распределение (смещен на 1) с вероятностью успеха $1/n$ и функцией вероятности

$$\mathbb{P}(N_n(r) = j) = \frac{\Gamma(j+r-1)}{(j-1)!\Gamma(r)} \left(\frac{1}{n}\right)^r \left(1 - \frac{1}{n}\right)^{j-1}, \\ r > 0,\;\; j = 1, 2, \ldots \quad (6)$$

Теперь получим разложение Чебышёва–Эджворта для нормированной статистики типа Хотеллинга. Функция F-распределения $F(x; a, b)$ — это абсолютно непрерывная функция распределения вероятности, заданная плотностью

$$f(x; a, b) = \\ = \frac{1}{B(a/2, b/2)} \left(\frac{a}{b}\right)^{a/2} x^{a/2-1} \left(1 + \frac{a}{b}x\right)^{-(a+b)/2}, \\ x > 0\,.$$

Лемма 1. *Пусть $r > 1$, случайная величина $N_n(r)$ определена формулой* (6)*, тогда*

$$\mathbb{E}\left(N_n(r)\right)^{-2} \le C(r) \begin{cases} n^{-r}, & 1 < r < 2\,; \\ \ln(n) n^{-2}, & r = 2\,; \\ n^{-2}, & r > 2\,. \end{cases} \quad (7)$$

В случае если $r = 2$, скорость сходимости в (7)*, не может быть улучшена.*

Используя теорему 1 из работы [4], получаем следующий результат.

Теорема 3. *Пусть статистика T_m определена формулой* (1)*. Пусть также дискретная случайная величина $N_n = N_n(r)$ с параметром $r > 1$ имеет распределение, задаваемое* (6)*, и независима от $W_p(q, I_p)$ и $W_p(n, I_p)$. Рассмотрим статистику $T_{N_n} = g(n)\,\mathrm{tr} S_h S_{N_n}^{-1}$. Асимптотическое разложение для случайного обьема $N_n(r)$ с $r > 1$ из* [4, теорема 1] *справедливо с $g(n) = \mathbb{E}\left(N_n(r)\right) = r(n-1) + 1$. Тогда существует константа $C = C(r) > 0$ такая, что для всех $n \in \mathbf{N}$*

$$\sup_x \left|\mathbb{P}\left(g(n)\,\mathrm{tr} S_h S_{N_n}^{-1} \le x\right) - F_{2;n}(x)\right| \le$$
$$\le C \begin{cases} n^{-r}, & 1 < r < 2\,; \\ \ln(n)\, n^{-2}, & r = 2\,; \\ n^{-2}, & r > 2\,; \end{cases} \qquad (8)$$

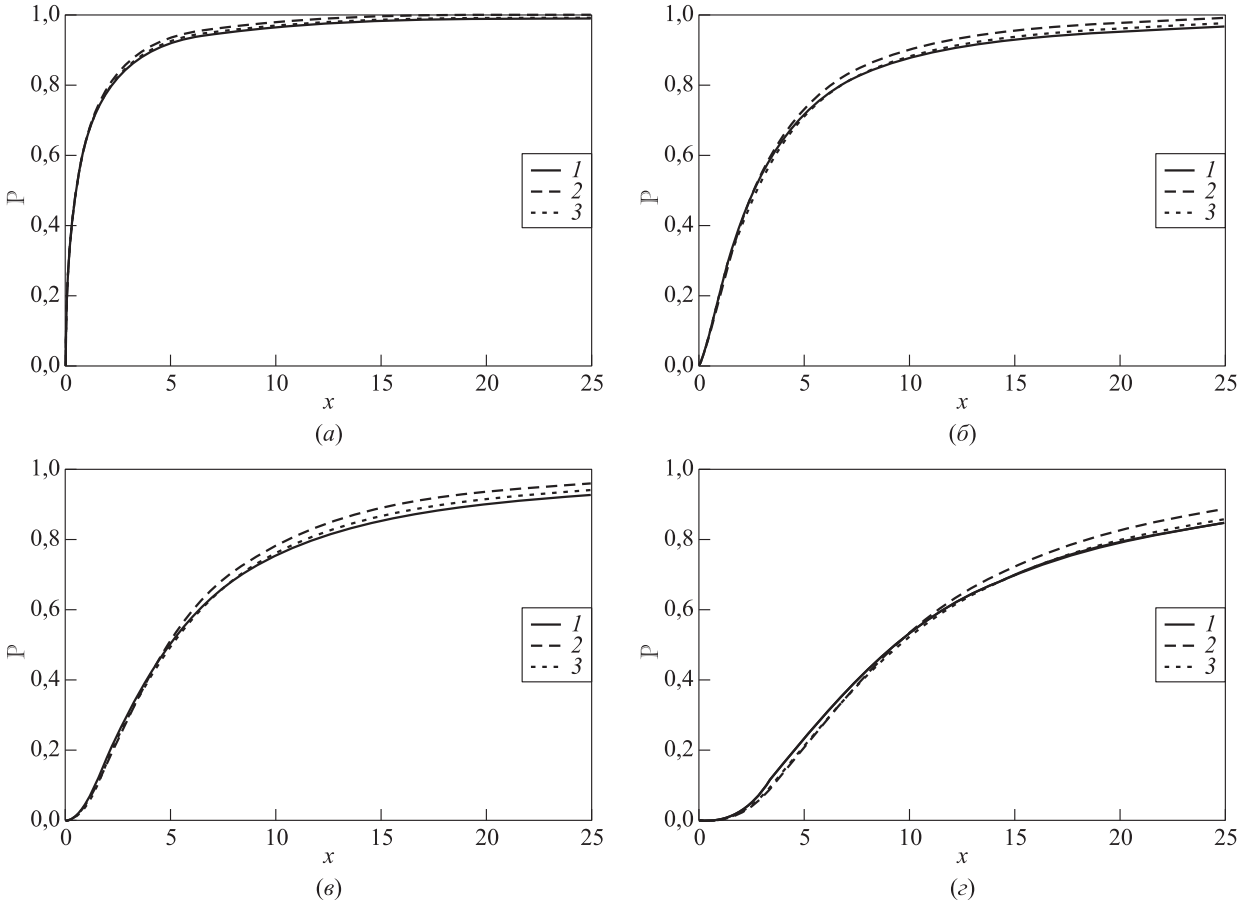
где

$$F_{2;n}(x) = F\left(\frac{x}{k}; k, 2r\right) + \frac{1}{n}\frac{(r-2)x}{2rk}\left(f\left(\frac{x}{k}; k, 2r\right) - f\left(\frac{r-1}{rk}x; k, 2r-2\right)\right) + \frac{k}{4(r(n-1)+1)} \times \sum_{j=0}^{2} a_j F\left(\frac{r-1}{(k+2j)r}x; k+2j, 2r-2\right) + \frac{k}{4n(r(n-1)+1)} \times \sum_{j=0}^{2} a_j \left[\frac{2-r}{2(r-1)} F\left(\frac{r-1}{r(k+2j)}x; k+2j, 2r-2\right) + \frac{r}{2(r-1)} F\left(\frac{r-2}{r(k+2j)}x; k+2j, 2r-4\right) - \frac{(2-r)x}{2r(k+2j)} f\left(\frac{r-1}{r(k+2j)}x; k+2j, 2r-2\right) - \frac{(r-2)x}{2(k+2j)(r-1)} \times f\left(\frac{r-2}{r(k+2j)}x; k+2j, 2r-4\right)\right].$$

На рисунке демонстрируется преимущество разложения Чебышёва–Эджворта второго порядка над разложением первого порядка в приближении эмпирической функции распределения.

4 Разложение Корниша–Фишера

Рассмотрим частный случай основного результата теоремы 3. Пусть в (8) $r = 3/2$. Обозначим $\mathbf{P}\left(g(n)\,\mathrm{tr}\, S_h S_{N_n}^{-1} \le x\right) = \bar{F}(x)$. Тогда



Эмпирическая функция распределения $\mathbb{P}\left(g(n)\,\mathrm{tr}\, S_h S_{N_n}^{-1} \le x\right)$ (*1*), аппроксимация первого порядка $F\left(x/k; k, 2r\right)$ (*2*) и аппроксимация второго порядка $F_{2;n}(x)$ (*3*) при $p = 1$, $n = 10$ и $r = 3$: (*а*) $q = 1$; (*б*) 3; (*в*) 5; (*г*) $q = 9$

$$\bar{F}(x) = F_{2;n}(x)|_{r=3/2} + \mathcal{O}(1/n^{3/2}),\ \ n \to \infty\,, \quad (9)$$

где

$$F_{2;n}(x)|_{r=3/2} = F\left(\frac{x}{k};k,3\right) - \frac{1}{n}\,\frac{x}{6k}\,f\left(\frac{x}{k};k,3\right) + \\ + \frac{1}{n}\,\frac{x}{6k}\,f\left(\frac{x}{3k};k,1\right) + \\ + \frac{k}{2(3n-1)}\sum_{j=0}^{2} a_j F\left(\frac{1}{(k+2j)3}\,x;k+2j,1\right).$$

Домножим и разделим третий и четвертый член разложения на плотность $f\,(x/k;k,3)$:

$$F_{2;n}(x)|_{r=3/2} = \\ = F\left(\frac{x}{k};k,3\right) + \frac{1}{n}\,d_2(x) f\left(\frac{x}{k};k,3\right), \quad (10)$$

где

$$d_2(x) = -\frac{x}{6k} + \frac{x}{6k}\,\frac{f\,(x/(3k);k,1)}{f\,(x/k;k,3)} + \\ + \frac{k}{2(3-1/n)}\sum_{j=0}^{2} a_j \frac{F\,(x/((k+2j)3);k+2j,1)}{f\,(x/k;k,3)}\,.$$

Перепишем исходное выражение (9), используя (10):

$$\bar{F}(x) = F\left(\frac{x}{k};k,3\right) + \frac{1}{n}\,d_2(x) f\left(\frac{x}{k};k,3\right) + \\ + \mathcal{O}\left(\frac{1}{n^{3/2}}\right), \quad n \to \infty\,. \quad (11)$$

Используя разложение Чебышёва–Эджворта (11) [4, утверждение 2], вытекающее из более общих утверждений (см., например, работы [7, гл. 5.6.1] и[8]), с $a_1(x) = 0$, $a_2(x) = d_2(x)$, $g(x) = f(x/k;k,3)$, $G(x) = F(x/k;k,3)$, получаем следующую теорему.

Теорема 4. *В условиях теоремы* 3 *пусть* $x = x_\alpha$, $u = u_\alpha$ *—* α*-квантили нормированной статистики* $\mathbf{P}\left(g(n)\,\mathrm{tr}\,S_h S_{N_n}^{-1} \le x\right)$ *и предельного* F*-распределения соответственно. Тогда справедливо следующее асимптотическое разложение для* $n \to \infty$:

$$x = ku + \frac{u}{6}\left(1 - \frac{f\,(u/3;k,1)}{f\,(u;k,3)}\right) n^{-1} - \\ = \frac{k}{6}\sum_{j=0}^{2} a_j \frac{F\,(uk/((k+2j)3);k+2j,1)}{f\,(u;k,3)}\,n^{-1} + \\ + \mathcal{O}(n^{-3/2}).$$

5 Доказательства

Доказательство теоремы 2.

Используя формулу полной вероятности, получаем

$$\mathbb{P}\left(g(n)\,\mathrm{tr}\,S_h S_{N_n}^{-1} \le x\right) = \\ = \mathbb{E}\mathbb{P}\left(N_n\,\mathrm{tr}\,S_h S_{N_n}^{-1} \le \frac{N_n}{g(n)}\,x | N_n\right) = \\ = \sum_{l=1}^{\infty} \mathbb{P}\left(l\,\mathrm{tr}\,S_h S_e^{-1} \le \frac{l}{g(n)}\,x\right)\mathbb{P}(N_n = l)\,. \quad (12)$$

Далее введем $F_n(x)$, подставив оценку (3) для статистики Хотеллинга из теоремы 1 без остаточного члена и оценку (4) без остаточного члена для нормированной функции распределения размера выборки из условия 1 в (12):

$$F_n(x) = \\ = \mathbb{P}\left(g(n)\,\mathrm{tr}\,S_h S_{N_n}^{-1} \le x\right) \overset{(3)}{=} \mathbb{E}\left(G_k\left(\frac{N_n}{g(n)}\,x\right) + \right. \\ \left. + \frac{k}{4N_n}\sum_{j=0}^{2} a_j G_{k+2j}\left(\frac{N_n}{g(n)}\,x\right)\right) = \int\limits_{1/g(n)}^{\infty}\left(G_k\,(yx) + \right. \\ \left. + \frac{k}{4g(n)y}\sum_{j=0}^{2} a_j G_{k+2j}\,(yx)\right) d\mathbb{P}\left(\frac{N_n}{g(n)} < y\right) \overset{(4)}{=} \\ \overset{(4)}{=} \int\limits_{1/g(n)}^{\infty}\left(G_k\,(yx) + \frac{k}{4g(n)y} \times \right. \\ \left. \times \sum_{j=0}^{2} a_j G_{k+2j}\,(yx)\right) d\left(H(y) - \sum_{i=1}^{m}\frac{1}{n^{i/2}}\,h_i(y)\right) = \\ = \int\limits_{1/g(n)}^{\infty} G_k(xy)\,dH(y) + \\ + \sum_{i=1}^{m}\frac{1}{n^{i/2}}\int\limits_{1/g(n)}^{\infty} G_k(xy)\,dh_i(y) + \\ + \frac{k}{4g(n)}\int\limits_{1/g(n)}^{\infty}\sum_{j=0}^{2}\frac{a_j}{y}\,G_{k+2j}(xy)\,dH(y) + \frac{k}{4g(n)} \times \\ \times \sum_{i=1}^{m}\frac{1}{n^{i/2}}\int\limits_{1/g(n)}^{\infty}\frac{1}{y}\sum_{j=0}^{2} a_j G_{k+2j}(xy)\,dh_i(y)\,. \quad (13)$$

Найдем теперь оценки для $\sup_x \left|\mathbb{P}\left(g(n)\,\mathrm{tr}\,S_h S_{N_n}^{-1} \le \right.\right. \\ \left.\left. \le x\right) - F_n(x)\right|$. Введем обозначение

$$\psi(n;y) = \mathbb{P}\left(\frac{N_n}{g(n)} < y\right) - H(y) - \sum_{i=1}^{m}\frac{1}{n^{i/2}}\,h_i(y)\,.$$

Согласно (13),

$$\sum_{l=1}^{\infty}\left(G_k\left(\frac{l}{g(n)}\,x\right)+\right.$$
$$\left.+\frac{k}{4l}\sum_{j=0}^{2}a_jG_{k+2j}\left(\frac{l}{g(n)}x\right)\right)\mathbb{P}(N_n=l)=$$
$$=\int\limits_{1/g(n)}^{\infty}\left(G_k\left(yx\right)+\right.$$
$$\left.+\frac{k}{4g(n)y}\sum_{j=0}^{2}a_jG_{k+2j}\left(yx\right)\right)d\mathbb{P}\left(\frac{N_n}{g(n)}<y\right),$$

поэтому

$$\sup_x\left|\mathbb{P}\left(g(n)\operatorname{tr}S_hS_{N_n}^{-1}\le x\right)-F_n(x)\right|\le I_{1n}+I_{n2},$$

где

$$I_{1n}=\sup_x\left|\int\limits_{1/g(n)}^{\infty}\left(G_k\left(yx\right)+\right.\right.$$
$$\left.\left.+\frac{k}{4g(n)y}\sum_{j=0}^{2}a_jG_{k+2j}(yx)\right)d\psi(n;y)\right|;$$
$$I_{2n}=\sum_{l=1}^{\infty}\sup_x\left|\mathbb{P}\left(l\operatorname{tr}S_hS_l^{-1}\le\frac{l}{g(n)}x\right)-\right.$$
$$\left.-G_k\left(\frac{l}{g(n)}\,x\right)-\frac{k}{4l}\sum_{j=0}^{2}a_jG_{k+2j}\left(\frac{l}{g(n)}\,x\right)\right|\times$$
$$\times\,\mathbb{P}(N_n=l).$$

Используя формулу интегрирования по частям и условие 1, получаем, что существует константа $C_3>0$ такая, что

$$I_{1n}\le\frac{C_3}{n^\beta}+\sup_x\int\limits_{1/g(n)}^{\infty}|(\psi(n;y)|\left|\frac{\partial}{\partial y}\left(G_k\left(yx\right)+\right.\right.$$
$$\left.\left.+\frac{k}{4g(n)y}\sum_{j=0}^{2}a_jG_{k+2j}\left(yx\right)\right)\right|dy\le\frac{C_3}{n^\beta}+\frac{C_2M_n}{n^\beta}\,.$$

Используя результат теоремы 1, получаем

$$I_{2n}\le\sum_{l=1}^{\infty}\frac{C_1}{l^2}\,\mathbb{P}(N_n=l)=C_1\mathbb{E}N_n^{-2}.\ \square$$

Доказательство следствия 1 и леммы 1 аналогичны доказательствам утверждения 1 и леммы 1 из работы [4].

Доказательство теоремы 3.

Обозначим

$$J_1(x)=\int\limits_0^{\infty}G_k(xy)dG_{r,r}(y)\,;$$
$$J_2(x)=$$
$$=\int\limits_0^{\infty}G_k(xy)\,d\left[\frac{(y-1)(2-r)+2Q_1\left(g(n)y\right)}{2r}\,g_{r,r}(y)\right];$$
$$J_3(x)=\int\limits_0^{\infty}\sum_{j=0}^{2}\frac{a_j}{y}\,G_{k+2j}(xy)\,dG_{r,r}(y)\,;$$
$$J_4(x)=\int\limits_0^{\infty}\frac{1}{y}\sum_{j=0}^{2}a_j\times$$
$$\times\,G_{k+2j}(xy)\,d\left[\frac{(y-1)(2-r)+2Q_1(g(n)y)}{2r}\,g_{r,r}(y)\right].$$

Тогда общий вид функции (5) из следствия 1 запишется в виде:

$$F_{2;n}(x)=J_1(x)+\frac{1}{n}\,J_2(x)+$$
$$+\frac{k}{4g(n)}\,J_3(x)+\frac{k}{4ng(n)}\,J_4(x).$$

Рассмотрим интеграл $J_1(x)$:

$$\frac{\partial}{\partial x}\,J_1(x)=\int\limits_0^{\infty}yg_k(xy)g_{r,r}(y)\,dy=$$
$$=\frac{r^r\,x^{k/2-1}}{\Gamma(r)\,\Gamma(k/2)2^{k/2}}\int\limits_0^{\infty}y^{r+k/2-1}e^{-(r+x/2)y}dy\,.$$

Используя формулу 2.3.3.1 из [9, с. 259]

$$\int\limits_0^{\infty}x^{\alpha-1}e^{-px}\,dx=\Gamma(\alpha)\,p^{-\alpha},\ \ \alpha,\,p>0\,,$$

с $\alpha=r+k/2$ и $p=r+x/2$, получаем

$$\frac{\partial}{\partial x}\,J_1(x)=\frac{r^r}{B(k/2,r)2^{k/2}}\,\frac{x^{k/2-1}}{(x/2+r)^{k/2+r}}\,,\ \ x>0\,.$$

Тогда

$$J_1(x)=\int\limits_0^{x}\frac{r^r}{B(k/2,r)2^{k/2}}\,\frac{t^{k/2-1}dt}{(t/2+r)^{k/2+r}}=$$
$$=\frac{1}{B(k/2,r)(2r)^{k/2}}\int\limits_0^{x}\frac{t^{k/2-1}dt}{(t/(2r)+1)^{k/2+r}}=$$
$$=\left\{y=\frac{t}{k},\ \ dt=k\,dy\right\}=F\left(\frac{x}{k};k,2r\right).\quad(14)$$

Обозначим

$$J_{3_j}(x) = \int\limits_0^\infty \frac{1}{y}\, G_{k+2j}(xy)\, dG_{r,r}(y),\ j = 0, 1, 2;$$

$$J_3(x) = \sum_{j=0}^{2} a_j J_{3_j}(x).$$

По аналогии с $(\partial/\partial x)J_1$ для $(\partial/\partial x)J_{3_0}$ получаем

$$\frac{\partial}{\partial x}\, J_{3_0}(x) = \\ = \frac{r\Gamma(k/2+r-1)}{\Gamma(k/2)\Gamma(r)(2r)^{k/2}}\, x^{k/2-1} \left(1 + \frac{x}{2r}\right)^{-(k/2+r-1)}.$$

Тогда

$$J_{3_0}(x) = \int\limits_0^x \frac{r\Gamma(k/2+r-1)}{\Gamma(k/2)(r-1)\Gamma(r-1)(2r)^{k/2}} \times \\ \times t^{k/2-1} \left(1 + \frac{t}{2r}\right)^{-(k/2+r-1)} dt = \\ = \left\{ y = \frac{r-1}{rk}\, t,\ \ dt = \frac{kr}{r-1}\, dy \right\} = \\ = \frac{r}{r-1} F\left(\frac{r-1}{rk}\, x; k, 2r-2\right).$$

Аналогично получаем

$$J_{3_1}(x) = \frac{r}{r-1} F\left(\frac{r-1}{r(k+2)} x; k+2, 2r-2\right);$$

$$J_{3_2}(x) = \frac{r}{r-1} F\left(\frac{r-1}{r(k+4)} x; k+4, 2r-2\right)$$

и, таким образом,

$$J_3(x) = \\ = \frac{r}{r-1} \sum_{j=0}^{2} a_j F\left(\frac{r-1}{r(k+2j)}\, x; k+2j, 2r-2\right). \quad (15)$$

Для вычисления $J_2(x)$ используем интегрирование по частям:

$$J_2(x) = -x \times \\ \times \int\limits_0^\infty g_k(xy) \frac{(y-1)(2-r) + 2Q_1(g(n)y)}{2r} g_{r,r}(y)\, dy = \\ = -\frac{x(2-r)}{2r}\, J_{2_1}(x) + \frac{x(2-r)}{2r}\, J_{2_2}(x) - \\ - \frac{x}{r}\, J_{2_3}(x), \quad (16)$$

где

$$J_{2_1}(x) = \int\limits_0^\infty y g_k(xy) g_{r,r}(y)\, dy;$$

$$J_{2_2}(x) = \int\limits_0^\infty g_k(xy) g_{r,r}(y)\, dy;$$

$$J_{2_3}(x) = \int\limits_0^\infty g_k(xy) g_{r,r}(y) Q_1(g(n)y)\, dy.$$

Заметим, что

$$J_{2_1}(x) = \frac{\partial}{\partial x}\, J_1(x);\ \ J_{2_2}(x) = \frac{\partial}{\partial x}\, J_{3_0}(x).$$

Рассмотрим третье слагаемое $J_{2_3}(x)$:

$$J_{2_3}(x) = \frac{r^r\, x^{k/2-1}}{\Gamma(r)\Gamma(k/2) 2^{k/2}} \times \\ \times \int\limits_0^\infty y^{r+k/2-2} e^{-(r+x/2)y} Q_1(g(n)y)\, dy\,.$$

Применяя технику из доказательства теоремы 2 из работы [4], для $J_4^*(x)$ получаем:

$$n^{-1}\ |J_{2_3}| \le \frac{c(r,k)}{n^r} \sum_{k=1}^{\infty} k^{-r} = \frac{c_1(r,k)}{n^r}\,.$$

Подставив выражения для $J_{2_1}(x)$ и $J_{2_2}(x)$ в (16), получаем

$$J_2(x) = \frac{(r-2)x}{2rk} \left(f\left(\frac{x}{k}; k, 2r\right) - \right. \\ \left. - f\left(\frac{r-1}{rk}\, x; k, 2r-2\right)\right). \quad (17)$$

Обозначим

$$J_{4_j}(x) = \int\limits_0^\infty \frac{1}{y}\, G_{k+2j}(xy)\, dh_2(y)\,,\ \ j = 0, 1, 2\,;$$

$$J_4(x) = \sum_{j=0}^{2} a_j J_{4_j}(x)$$

и рассмотрим J_{4_0}. Используя интегрирование по частям, получаем

$$J_{4_0}(x) = -\int\limits_0^\infty \left(-\frac{1}{y^2}\, G_k(xy) + \frac{x}{y}\, g_k(xy)\right) h_2(y)\, dy = \\ = \int\limits_0^\infty \frac{1}{y^2}\, G_k(xy) h_2(y)\, dy - x \int\limits_0^\infty \frac{1}{y}\, g_k(xy) h_2(y)\, dy = \\ = J'(x) - xJ''(x). \quad (18)$$

Заметим, что $J''(x) = (\partial/\partial x)J'(x)$, поэтому достаточно рассмотреть $J''(x)$:

$$J''(x) = \int_0^\infty \frac{1}{y}\, g_k(xy) \times \\ \times \frac{(y-1)(2-r) + 2Q_1(g(n)y)}{2r}\, g_{r,r}(y)\, dy + \\ + \frac{1}{r}\int_0^\infty \frac{1}{y}\, g_k(xy) g_{r,r}(y) Q_1(g(n)y)\, dy = \\ = \frac{2-r}{2r}\, J_1''(x) - \frac{2-r}{2r}\, J_2''(x) + \frac{1}{r}\, J_3''(x).$$

Заметим, что

$$J_1''(x) = \frac{1}{k}\, f\left(\frac{r-1}{rk}\, x; k, 2r-2\right)$$

и оценка для $J_3''(x)$ строится аналогично оценке для J_{2_3}, но с $\alpha = r + k/2 - 2$. Рассмотрим $J_2''(x)$:

$$J_2''(x) = \\ = \frac{r^r\, x^{k/2-1}}{\Gamma(r)\Gamma(k/2)2^{k/2}} \int_0^\infty y^{r+k/2-3} e^{-(r+x/2)y}\, dy = \\ = \left\{x = \frac{rk}{r-2}t\right\} = \frac{r}{(r-1)k}\, f\left(\frac{r-2}{rk}\, x; k, 2r-4\right),$$

откуда получаем

$$J''(x) = \frac{2-r}{2rk}\, f\left(\frac{r-1}{rk}\, x; k, 2r-2\right) + \\ + \frac{r-2}{2k(r-1)}\, f\left(\frac{r-2}{rk}\, x; k, 2r-4\right); \quad (19)$$

$$J'(x) = \frac{2-r}{2(r-1)}\, F\left(\frac{r-1}{rk}\, x; k, 2r-2\right) + \\ + \frac{r}{2(r-1)}\, F\left(\frac{r-2}{rk}\, x; k, 2r-4\right). \quad (20)$$

Подставляя (19) и (20) в (18), получаем

$$J_{4_0}(x) = \frac{2-r}{2(r-1)}\, F\left(\frac{r-1}{rk}\, x; k, 2r-2\right) + \\ + \frac{r}{2(r-1)}\, F\left(\frac{r-2}{rk}\, x; k, 2r-4\right) - \\ - \frac{(2-r)x}{2rk}\, f\left(\frac{r-1}{rk}\, x; k, 2r-2\right) - \\ - \frac{(r-2)x}{2k(r-1)}\, f\left(\frac{r-2}{rk}\, x; k, 2r-4\right).$$

Аналогично получаем, что

$$J_{4_1}(x) = \frac{2-r}{2(r-1)}\, F\left(\frac{r-1}{r(k+2)}\, x; k+2, 2r-2\right) + \\ + \frac{r}{2(r-1)}\, F\left(\frac{r-2}{r(k+2)}\, x; k+2, 2r-4\right) - \\ - \frac{(2-r)x}{2r(k+2)}\, f\left(\frac{r-1}{r(k+2)}\, x; k+2, 2r-2\right) - \\ - \frac{(r-2)x}{2(k+2)(r-1)}\, f\left(\frac{r-2}{r(k+2)}\, x; k+2, 2r-4\right);$$

$$J_{4_2}(x) = \frac{2-r}{2(r-1)}\, F\left(\frac{r-1}{r(k+4)}\, x; k+4, 2r-2\right) + \\ + \frac{r}{2(r-1)}\, F\left(\frac{r-2}{r(k+4)}\, x; k+4, 2r-4\right) - \\ - \frac{(2-r)x}{2r(k+4)}\, f\left(\frac{r-1}{r(k+4)}\, x; k+4, 2r-2\right) - \\ - \frac{(r-2)x}{2(k+4)(r-1)}\, f\left(\frac{r-2}{r(k+4)}\, x; k+4, 2r-4\right).$$

Таким образом,

$$J_4(x) = \\ = \sum_{j=0}^{2} a_j \left[\frac{2-r}{2(r-1)}\, F\left(\frac{r-1}{r(k+2j)}\, x; k+2j, 2r-2\right) + \right. \\ + \frac{r}{2(r-1)}\, F\left(\frac{r-2}{r(k+2j)} x; k+2j, 2r-4\right) - \\ - \frac{(2-r)x}{2r(k+2j)}\, f\left(\frac{r-1}{r(k+2j)}\, x; k+2j, 2r-2\right) - \\ - \frac{(r-2)x}{2(k+2j)(r-1)} \times \\ \left. \times f\left(\frac{r-2}{r(k+2j)}\, x; k+2j, 2r-4\right)\right]. \quad (21)$$

Объединяя $|1/g(n) - 1/(rn)| \le \max\{2, r\}(r-1)(rn)^{-2}$, (14), (17), (15), (21) и лемму 1, получаем доказательство оценки (8).

6 Заключение

Доказанный в данной работе аналог теоремы переноса позволяет обобщить результаты работ [3, 4] для случая, когда статистика имеет распределения типа Хотеллинга случайного размера. Полученное в работе асимптотическое разложение типа Чебышёва–Эджворта для функции распределения вышеупомянутой статистики позволяет построить разложения типа Корниша–Фишера для данной статистики.

Автор выражает благодарность В. В. Ульянову и Г. Кристофу за полезные обсуждения задачи.

Литература

1. *Бенинг В. Е., Галиева Н. К., Королев В. Ю.* Асимптотические разложения для функций распределения статистик, построенных по выборкам случайного объема // Информатика и её применения, 2013. Т. 7. Вып. 2. С. 75–83.
2. *Бенинг В. Е., Галиева Н. К., Королев В. Ю.* Оценки скорости сходимости для функций распределения асимптотически нормальных статистик, основанных на выборках случайного объема // Вестник Тверского гос. ун-та. Сер.: Прикладная математика, 2012. Т. 17. С. 53–65.
3. *Марков А. С., Монахов М. М., Ульянов В. В.* Разложения типа Корниша–Фишера для распределений статистик, построенных по выборкам случайного размера // Информатика и её применения, 2016. Т. 10. Вып. 2. С. 84–91.
4. *Кристоф Г., Монахов М. М., Ульянов В. В.* Разложения Чебышева–Эджворта и Корниша–Фишера второго порядка для распределений статистик, построенных по выборкам случайного размера // Записки научных семинаров ПОМИ, 2017. Т. 466. С. 167–207.
5. *Ulyanov V. V., Aoshima M., Fujikoshi Y.* Non-asymptotic results for Cornish–Fisher expansions // J. Math. Sci., 2016. Vol. 218. No. 3. P. 363–368.
6. *Fujikoshi Y., Ulyanov V. V., Shimizu R.* L_1-norm error bounds for asymptotic expansions of multivariate scale mixtures and their applications to Hotelling's generalized T_0^2 // J. Multivariate Anal., 2005. Vol. 96. P. 1–19.
7. *Fujikoshi Y., Ulyanov V. V., Shimizu R.* Multivariate statistics: High-dimensional and large-sample approximations. — Wiley ser. in probability and statistics. – Hoboken, NJ, USA: Wiley, 2010. 568 p.
8. *Ulyanov V. V.* Cornish–Fisher expansions // International encyclopedia of statistical science / Ed. M. Lovric. — Berlin: Springer, 2011. P. 312–315.
9. *Прудников А. П., Брычков Ю. А., Маричев О. И.* Интегралы и ряды. — М.: Наука, 1981. Т. 1. 259 с.

Поступила в редакцию 22.06.2020

CHEBYSHEV–EDGEWORTH EXPANSIONS FOR DISTRIBUTIONS OF GENERALISED HOTELLING-TYPE STATISTICS BASED ON RANDOM SIZE SAMPLES

M. M. Monakhov

Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation

Abstract: The general transfer theorem for the distribution function of asymptotically normal statistics was generalized on the Hotelling-type statistics case and analog of general transfer theorem for the distribution function of Hotelling-type statistics with random size was proved. It allowed to obtain the Chebyshev–Edgeworth expansion for initial Hotelling-type statistics. The explicit form of the Chebyshev–Edgeworth expansion was obtained for the case when the random sample size distribution is the negative binomial distribution shifted by 1. The limit distribution for this case was F-distribution. The Cornish–Fisher expansion was obtained for the special case of parameter of random sample size. The computational experiment was conducted and graphs were plotted for Chebyshev–Edgeworth expansion illustration.

Keywords: generalised Chebyshev–Edgeworth expansion; Cornish–Fisher expansion; sample with random size; F-disribution; Hotelling-type statstics

DOI: 10.14357/19922264210211

Acknowledgments

The research was conducted in accordance with the program of the Moscow Center for Fundamental and Applied Mathematics.

References

1. Bening, V. E., N. K. Galieva, and V. Yu. Korolev. 2013. Asimptoticheskie razlozheniya dlya funktsiy raspredeleniya statistik, postroennykh po vyborkam sluchaynogo ob"ema [Asymptotic expansions for the distribution functions of statistics constructed from samples with random sizes]. *Informatika i ee Primeneniya — Inform. Appl.* 7(2):75–83.
2. Bening, V. E., N. K. Galieva, and V. Yu. Korolev. 2012. Otsenki skorosti skhodimosti dlya funktsiy raspredeleniya

asimptoticheski normal'nykh statistik, osnovannykh na vyborkakh sluchaynogo ob"ema [On rate of convergence in distribution of asymptotically normal statistics based on samples of random size]. *Vestnik Tverskogo gos. un-ta. Ser. Prikladnaya matematika* [Bull. of the Tverskoy State University. Ser. Appl. Math.] 17:53–65.

3. Markov, A. S., M. M. Monakhov, and V. V. Ulyanov. 2016. Razlozheniya tipa Kornisha–Fishera dlya raspredeleniy statistik, postroennykh po vyborkam sluchaynogo razmera [Generalized Cornish–Fisher expansions for distributions of statistics based on samples of random size]. *Informatika i ee Primeneniya — Inform. Appl.* 10(2):84–91.
4. Christoph, G., M. M. Monakhov, and V. V. Ulyanov. 2017. Razlozheniya Chebysheva–Edzhvorta i Kornisha–Fishera vtorogo poryadka dlya raspredeleniy statistik, postroyennykh po vyborkam sluchaynogo razmera [Second order Chebyshev–Edgeworth and Cornish–Fisher expansions for distributions of statistics constructed from samples with random sizes]. *Zapiski nauchnykh seminarov POMI* [POMI Notes of Scientific Seminars] 466:167–207.
5. Ulyanov, V. V., M. Aoshima, and Y. Fujikoshi. 2016. Nonasymptotic results for Cornish–Fisher expansions. *J. Math. Sci.* 218(3):363–368.
6. Fujikoshi, Y., V. V. Ulyanov, and R. Shimizu. 2005. L_1-norm error bounds for asymptotic expansions of multivariate scale mixtures and their applications to Hotelling's generalized T_0^2. *J. Multivariate Anal.* 96:1–19.
7. Fujikoshi, Y., V. V. Ulyanov, and R. Shimizu. 2010. *Multivariate statistics: High-dimensional and large-sample approximations*. Wiley ser. in probability and statistics. Hoboken, NJ: Wiley. 568 p.
8. Ulyanov, V. V. 2011. Cornish–Fisher expansions. *International encyclopedia of statistical science*. Ed. M. Lovric. Berlin: Springer. 312–315.
9. Prudnikov, A. P., Yu. A. Brychkov, and O. I. Marichev. 1992. *Integraly i ryady* [Integrals and series]. Moscow: Nauka. Vol. 1. 259 p.

Received June 22, 2020

Contributor

Monakhov Mikhail M. (b. 1993) — laboratory assistant, Moscow Center for Fundamental and Applied Mathematics, M. V. Lomonosov Moscow State University, 1-52 Leninskie Gory, GSP-1, Moscow 119991, Russian Federation; mih_monah@mail.ru

АЛГОРИТМЫ СЖАТИЯ ДАННЫХ МАССИВОВ СИЛОВЫХ КРИВЫХ I: КОДИРОВАНИЕ ОШИБОК ПРЕДСКАЗАНИЯ

Д. В. Сушко[1]

Аннотация: Рассмотрена задача обратимого (без потерь) сжатия данных массивов силовых кривых — трехмерных массивов, элементы которых суть 16-битные целые числа. Такие массивы являются результатом сканирования микрообъектов на атомно-силовом микроскопе (АСМ) в режиме измерения силовых карт. Предложены алгоритмы обратимого сжатия массивов силовых кривых, основанные на универсальном арифметическом кодировании ошибок их предсказания. Применены два метода универсального кодирования. Первый основан на использовании статистической модели источника с вычислимой последовательностью состояний и предполагает разложение всей последовательности ошибок предсказания на две независимо кодируемые подпоследовательности. Второй предполагает выбор подходящего веса при построении используемых в арифметическом кодировании кодовых вероятностей. Для предложенных алгоритмов на пяти тестовых массивах построены оценки скорости кодирования. Результаты показывают, что использование комбинации упомянутых выше методов универсального кодирования позволяет заметно уменьшить скорость кодирования. Скорости кодирования тестовых массивов наиболее эффективным алгоритмом среди предложенных практически применимых алгоритмов составили 3,9285, 3,5268, 3,5024, 4,2813 и 4,2246 бит/пиксель.

Ключевые слова: атомно-силовой микроскоп; массив силовых кривых; обратимое сжатие; арифметическое кодирование; универсальное кодирование

DOI: 10.14357/19922264210212

1 Введение

В настоящее время широко распространенным методом исследования микрообъектов (например, клеток, вирусов, белков, нуклеиновых кислот и др. в микробиологии) стало их сканирование на АСМ в режиме измерения силовых карт. Результатом таких исследований являются массивы силовых кривых – трехмерные массивы данных большого объема (в типичном случае $\sim$ 70 МБ). Необходимость долгосрочного хранения и передачи по каналам связи таких данных делает задачу их сжатия весьма актуальной, а поскольку получение данных связано с проведением трудоемкого и длительного эксперимента, потери при сжатии недопустимы, т. е. сжатие должно быть обратимым (без потерь).

Разработка алгоритмов обратимого сжатия данных массивов силовых кривых представляет интерес с теоретической точки зрения, так как распространенные алгоритмы обратимого сжатия ориентированы главным образом на сжатие данных одного из двух типов: текст или изображение, а массивы силовых кривых, очевидным образом, не относятся ни к одному из этих типов.

Исследование задачи обратимого сжатия данных массивов силовых кривых было начато в работе [1], где были определены потенциальные возможности некоторых алгоритмов сжатия. В качестве экспериментальных данных в работе были использованы массивы, полученные при сканировании мягких биологических образцов в режиме измерения силовых карт на микроскопе MultiMode V (Veeco, США). В число рассмотренных алгоритмов вошли стандартные алгоритмы (DEFLATE, JPEG 2000) и предложенные в [1] простые алгоритмы на основе арифметического кодирования. Были получены оценки скорости кодирования этих алгоритмов. Напомним, что *скоростью кодирования* R алгоритма называется отношение длины кодового слова L (в битах), порождаемого алгоритмом для описания массива данных, к полному числу элементов (пикселей) N этого массива; единица измерения скорости кодирования – бит/пиксель (бт/п). При этом коэффициент сжатия равен отношению длины элемента массива в битах к скорости кодирования.

Цель настоящей работы — предложить и исследовать более сложные алгоритмы обратимого сжатия, основанные на универсальном арифметическом кодировании ошибок предсказания. Скорость кодирования алгоритмов оценивается на тех же экспериментальных данных, что и в ра-

[1] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, dsushko@ipiran.ru

боте [1], что позволяет сравнивать получаемые результаты непосредственно. Вычисления проводятся программами, написанными на языке Python.

2 Массивы силовых кривых

Кратко рассмотрим вопросы, связанные с процессом сканирования и структурой массивов силовых кривых. Более подробное изложение приведено в [1], детальное описание технологии измерений с помощью АСМ и интерпретации соответствующих данных можно найти, например, в [2].

Принцип работы АСМ заключается в сканировании поверхности образца атомарно острой иглой (зондом), которая является частью гибкого кронштейна (кантилевера), закрепленного на пьезоэлектрическом двигателе. Силы, действующие на зонд со стороны поверхности, вызывают изгиб кантилевера, что приводит к перераспределению лазерного сигнала на фотодетекторе. Регистрируя величину этого сигнала и зная жесткость кантилевера, можно определить силу взаимодействия зонда с поверхностью.

Современные АСМ позволяют работать в режиме измерения силовых кривых и силовых карт. Силовая кривая представляет собой график зависимости силы взаимодействия зонда и поверхности образца от расстояния между ними. При фиксированном положении зонда в плоскости образца снимаются две кривые: кривая подвода (зонд приближается к образцу) и кривая отвода (зонд удаляется от образца). Пары силовых кривых снимаются для множества точек в поле наблюдения. В результате формируется карта силовых кривых — трехмерный массив данных.

Введем некоторые обозначения. Пусть $OXYZ$ — трехмерная декартова система координат. Твердая подложка образца располагается в горизонтальной плоскости OXY, поле наблюдения представляет собой прямоугольник в этой плоскости. Силовые кривые измеряются в узлах $(x, y) \equiv (x_i, y_j)$ равномерной прямоугольной решетки в поле наблюдения, $i = 0, 1, \ldots, I-1$, $j = 0, 1, \ldots, J-1$. Перебор узлов решетки осуществляется в следующем порядке: сначала по ширине (в направлении OY), затем по длине (в направлении OX). В каждом узле (x, y) решетки измеряется пара силовых кривых $F^{\mathrm{A}}_{(x,y)}(z)$ (кривая подвода) и $F^{\mathrm{R}}_{(x,y)}(z)$ (кривая отвода), $z \equiv z_k$, $k = 0, 1, \ldots, K-1$. Шаг по вертикали равномерный, высота z отсчитывается от поверхности образца в узле (x, y).

Значения элементов силовых кривых $F^{\mathrm{A,R}}_{(x,y)}(z)$ пропорциональны силе, действующей на зонд в точке пространства, находящейся на расстоянии z от поверхности образца и имеющей координаты (x, y) в поле наблюдения, в процессе подвода зонда к образцу и отвода зонда от образца. Значения F записываются в виде 16-битных целых чисел, т. е. целых чисел в диапазоне $[-2^{15}, 2^{15}-1]$. Увеличение значения F отвечает увеличению отталкивания (уменьшению притяжения) между зондом и поверхностью. При измерении кривой подвода $F^{\mathrm{A}}_{(x,y)}(z)$ зонд может достичь поверхности образца до того, как будет зарегистрировано необходимое число (K) значений. В таком случае осуществляется дополнение соответствующей строки до требуемой длины (K) минимально возможным значением -2^{15}, записываемым в конец строки.

Для всего массива силовых кривых будем использовать обозначение

$$\mathbf{V} = \{V(i,j,k)\},\ \ i = 0, 1, \ldots, I-1\,,\\ j = 0, 1, \ldots, J-1\,,\ \ k = 0, 1, \ldots, 2K-1\,.$$

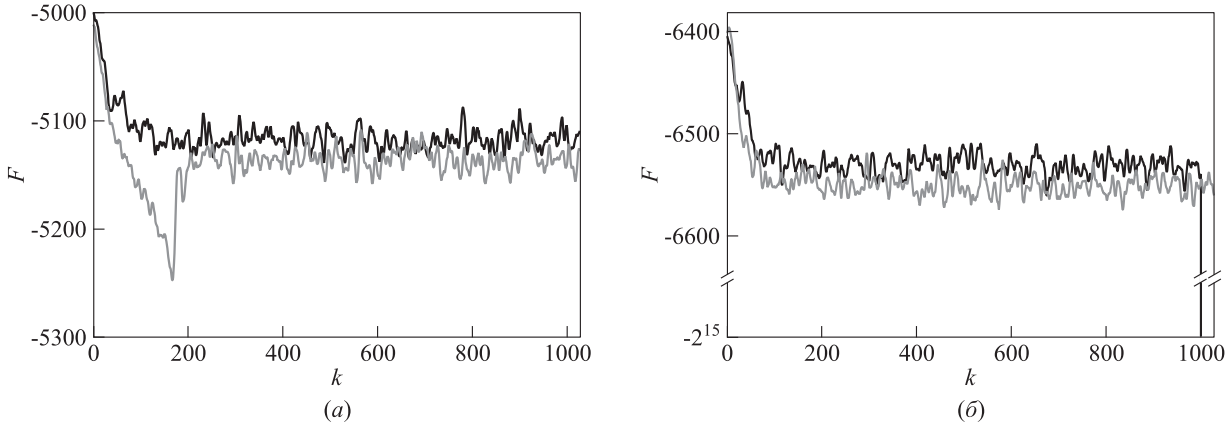
При этом

$$V(i,j,k) = \\ = \begin{cases} F^{\mathrm{A}}_{(x_i,y_j)}(z_k)\,, & k = 0, 1, \ldots, K-1\,; \\ F^{\mathrm{R}}_{(x_i,y_j)}(z_{k-K})\,, & k = K, K+1, \ldots, 2K-1\,. \end{cases}$$

В качестве экспериментальных данных используются пять массивов силовых кривых (I–V), полученных при сканировании образцов, представляющих собой абсорбированные из раствора на твердую подложку вирусы.

Первый образец (I) — это риновирус 2 на подложке из слюды, второй и третий образцы (II и III) — вирус мягкой мозаики ячменя на подложке из слюды, четвертый и пятый образцы (IV и V) — вирус табачной мозаики на подложке из стекла. Массивы силовых кривых имеют следующие размеры: $I = J = 64$, $K = 4096$ (массив I); $I = J = 128$, $K = 1024$ (массивы II–V). Полное число элементов всех массивов равно 2^{25}. Каждый элемент представляет собой 16-битное целое число.

В качестве иллюстрации на рисунке представлены пары силовых кривых образца II в двух разных узлах поля наблюдения. Кривые подвода изображены черным цветом, кривые отвода — серым. На фрейме (*б*) представлена кривая подвода, при измерении которой зонд достиг поверхности образца до того, как было зарегистрировано необходимое число ($K = 1024$) значений, и поэтому в конец строки было записано нужное число минимальных значений (-2^{15}).



Силовые кривые образца II

3 Алгоритмы кодирования ошибок предсказания

Общепринятый метод решения задач обратимого сжатия цифровых данных заключается в применении к исходному массиву данных некоторых обратимых преобразований, обеспечивающих декорреляцию его отсчетов и/или уменьшение диапазона их значений, и последующего арифметического кодирования [3] полученного таким образом массива как последовательности независимых отсчетов.

При арифметическом кодировании конечной числовой последовательности $\mathbf{x} = \{x_n\}$, $n = 0, 1, \ldots, N-1$, принимающей значения в априори известном диапазоне $\mathfrak{A}$, множество *условных кодовых распределений вероятностей* (или просто *кодовых распределений*) $\{q_n(a) = q_n(a|x_{n-1}, \ldots, x_0), a \in \mathfrak{A}\}$ используется для того, чтобы приписать последовательности $\mathbf{x}$ кодовую вероятность $Q(\mathbf{x})$ и кодовое слово (результат сжатия) длины

$$L(\mathbf{x}) = \left\lceil -\log_2 \left(\frac{Q(\mathbf{x})}{2} \right) \right\rceil = \left\lceil -\log_2 \prod_{n=0}^{N-1} q_n(x_n) + 1 \right\rceil \le \sum_{n=0}^{N-1} -\log_2 q_n(x_n) + 2\,. \quad (1)$$

Здесь $\lceil \cdot \rceil$ — результат округления вещественного числа до ближайшего целого вверх. При этом построение кодовых распределений, обеспечивающих получение возможно более коротких кодовых слов при неизвестной статистике, — задача универсального кодирования [4]. Отметим, что восстановление исходной последовательности по кодовому слову осуществляется в процессе декодирования без задержки. Это означает, что в момент восстановления очередного значения x_n декодеру уже известны все предыдущие значения $\{x_0, \ldots, x_{n-1}\}$ и кодовые распределения могут быть построены декодером так же, как они были ранее построены кодером в процессе кодирования. Это позволяет декодеру восстановить значение x_n.

В настоящей работе кодовые распределения строятся следующим образом:

$$q_n(a|x_{n-1}, \ldots, x_0) = \begin{cases} \dfrac{1 + w\theta_n(a)}{a_{\max}(\mathbf{x}) - a_{\min}(\mathbf{x}) + 1 + wn} & \text{при } a \in \\ & \in [a_{\min}(\mathbf{x}), a_{\max}(\mathbf{x})]\,; \\ 0 & \text{в противном случае}\,, \end{cases} \quad (2)$$

где $\theta_n(a)$ — число элементов, принимающих значение a, на начальном участке последовательности $\mathbf{x}$ до $(n-1)$-го члена включительно; $a_{\min}(\mathbf{x})$ и $a_{\max}(\mathbf{x})$ — нижняя и верхняя границы диапазона значений последовательности $\mathbf{x}$; параметр $w = 1, 2, \ldots$ — вес. В частном случае $w = 2$ формула (2) превращается в соответствующую формулу работы [1]. Границы $a_{\min}(\mathbf{x})$ и $a_{\max}(\mathbf{x})$ могут быть вычислены кодером и должны быть переданы декодеру помимо кодового слова, что, вообще говоря, несколько увеличивает скорость кодирования. Однако кодируемые в работе массивы имеют 2^{25} элементов, а для передачи значения одной границы требуется 2^5 бит; соответствующее увеличение скорости кодирования составляет 2^{-20} бт/п — величину, которой можно пренебречь.

В настоящей работе скорость кодирования оценивается по формуле

$$R(\mathbf{x}) \doteq \frac{1}{N} L(\mathbf{x}) = \frac{1}{N} \sum_{n=0}^{N-1} -\log_2 q_n(x_n)\,, \quad (3)$$

вытекающей из (1), если пренебречь бесконечно малым членом $2N^{-1}$.

Величины $\theta(x)/N$, $x \in \mathfrak{A}$, где $\theta(x)$ — число элементов, принимающих значение x, в последовательности $\mathbf{x}$, образуют *частотное* (или *эмпирическое*) распределение вероятностей значений последовательности. Величина

$$H(\mathbf{x}) = \sum_{x\in\mathfrak{A}} \frac{\theta(x)}{N}\left[-\log_2 \frac{\theta(x)}{N}\right] \qquad (4)$$

(используется соглашение о том, что $0 \cdot \log_2 0 = 0$) называется *квазиэнтропией* последовательности. Единица измерения квазиэнтропии — бт/п. Квазиэнтропия зависит только от самой последовательности и представляет собой нижнюю границу скорости арифметического кодирования (см., например, [4]). Разность $R - H \geq 0$ — *избыточность* арифметического кодирования. Величина избыточности характеризует качество решения одной из задач универсального кодирования — задачи построения кодовых распределений.

Арифметическое кодирование многомерного массива данных требует их одномерного упорядочения. В работе принято так называемое строчное упорядочение, при котором трехмерный массив $\mathbf{X}$ размерами (I, J, K) превращается в последовательность $\mathbf{x}$ размером $N = IJK$ так, что $x(n) = X(i, j, k)$, $n = JKi + Kj + k$.

В работе исследуются алгоритмы, использующие в качестве основного декорреляционного преобразования переход к ошибкам предсказания и два предваряющих его преобразования, которые увеличивают неравномерность распределения и уменьшают диапазон значений данных.

Первое преобразование заключается в разделении массива силовых кривых $\mathbf{V}$ на массивы кривых подвода $\mathbf{V}^{\mathrm{A}}$ и отвода $\mathbf{V}^{\mathrm{R}}$: $\mathbf{V} \to \{\mathbf{V}^{\mathrm{A}}, \mathbf{V}^{\mathrm{R}}\}$,

$$\begin{aligned} V^{\mathrm{A}}(i,j,k) &= V(i,j,k), \\ V^{\mathrm{R}}(i,j,k) &= V(i,j,K+k), \\ k &= 0,1,\ldots,K-1\,. \end{aligned} \qquad (5)$$

Второе преобразование применяется к полученному массиву кривых подвода $\mathbf{V}^{\mathrm{A}}$ и заключается в сужении диапазона значений этого массива, непомерно широкого из-за наличия значения -2^{15}, используемого для дополнения «неполных» строк: $\mathbf{V}^{\mathrm{A}} \to \bar{\mathbf{V}}^{\mathrm{A}}$,

$$\bar{V}^{\mathrm{A}} = \begin{cases} V^{\mathrm{A}}_{\mathrm{d\,min}} - 1, & V^{\mathrm{A}} = -2^{15}; \\ V^{\mathrm{A}}, & V^{\mathrm{A}} > -2^{15}, \end{cases} \qquad (6)$$

где $V^{\mathrm{A}}_{\mathrm{d\,min}}$ — динамический минимум значений массива $\mathbf{V}^{\mathrm{A}}$ (т. е. минимум значений без учета значения -2^{15}). Чтобы обратить преобразование сужения диапазона (6), декодеру нужна информация о том, встречалось ли значение -2^{15} в исходном массиве кривых подвода. Для передачи декодеру этой известной кодеру информации требуется один бит, и соответствующее увеличение скорости кодирования пренебрежимо мало.

Переход к ошибкам предсказания $\mathbf{X} \to \mathbf{D}$ для данного трехмерного массива $\mathbf{X} = \{X(i, j, k)\}$ размерами (I, J, K) имеет вид:

$$D(i,j,k) = \begin{cases} X(0,0,0), & i=j=k=0; \\ X(i,0,0) - X(i-1,0,0), & i=1,\ldots,I-1, \\ & j=k=0; \\ X(i,j,0) - X(i,j-1,0), & i=1,\ldots,I-1, \\ & j=1,\ldots,J-1,\ k=0; \\ X(i,j,k) - X(i,j,k-1), & i=1,\ldots,I-1, \\ & j=1,\ldots,J-1,\ k=1,\ldots,K-1. \end{cases} \qquad (7)$$

Преобразование применяется к каждому из двух полученных в результате предварительной обработки массивов: $\bar{\mathbf{V}}^{\mathrm{A}} \to \mathbf{D}^{\mathrm{A}}$; $\mathbf{V}^{\mathrm{R}} \to \mathbf{D}^{\mathrm{R}}$.

Все исследуемые в работе алгоритмы используют преобразования (5)–(7) и различаются применяемыми методами универсального кодирования. Отметим, что использование этих преобразований обеспечило наименьшую скорость кодирования массивов силовых кривых среди рассмотренных в [1] алгоритмов, основанных на кодировании ошибок предсказания.

Начнем с рассмотренного в [1] алгоритма A[1|0], который осуществляет независимое арифметическое кодирование массивов $\mathbf{D}^{\mathrm{A}}$ и $\mathbf{D}^{\mathrm{R}}$ с использованием построенных по формуле (2) с весом $w = 2$ кодовых распределений. В строке 1 табл. 1 приведена квазиэнтропия

$$H[1] \doteq H(\{\mathbf{D}^{\mathrm{A}}, \mathbf{D}^{\mathrm{R}}\}) = \frac{1}{2}\left[H(\mathbf{D}^{\mathrm{A}}) + H(\mathbf{D}^{\mathrm{A}})\right]$$

ошибок предсказания для массивов I–V, а в строке 4 — скорость кодирования

$$R[1|0] \doteq R(\{\mathbf{D}^{\mathrm{A}}, \mathbf{D}^{\mathrm{R}}\}) = \frac{1}{2}\left[R(\mathbf{D}^{\mathrm{A}}) + R(\mathbf{D}^{\mathrm{A}})\right]$$

алгоритма. Значения квазиэнтропии и скорости кодирования в табл. 1 приводятся в единицах бт/п с точностью до четырех знаков после десятичной запятой. Избыточность кодирования алгоритма составляет 0,0009–0,0022 бт/п в зависимости от массива.

Рассмотрим метод сжатия, основанный на использовании статистической модели *источника*

Таблица 1 Квазиэнтропия и скорость кодирования

№	Величина	I	II	III	IV	V
1	$H[1]$	3,9496	3,6922	3,6945	4,3091	4,2454
2	$H[2,\mathrm{opt}]$	3,9256	3,5261	3,5015	4,2806	4,2239
3	$H[2,\mathrm{fix}]$	3,9281	3,5261	3,5015	4,2806	4,2239
4	$R[1\|0]$	3,9504	3,6944	3,6959	4,3103	4,2463
5	$R[1\|\mathrm{opt}]$	3,9498	3,6927	3,6951	4,3096	4,2458
6	$R[1\|\mathrm{fix}]$	3,9498	3,6928	3,6951	4,3096	4,2458
7	$R[2,\mathrm{fix}\|0]$	3,9294	3,5288	3,5038	4,2823	4,2256
8	$R[2,\mathrm{fix}\|\mathrm{opt}]$	3,9285	3,5268	3,5024	4,2813	4,2246
9	$R[2,\mathrm{fix}\|\mathrm{fix}]$	3,9285	3,5268	3,5024	4,2813	4,2246

с вычислимой последовательностью состояний (см., например, [4]). Модель предполагает, что элементы последовательности $\mathbf{x}$ (в данном случае одномерно упорядоченные ошибки предсказания $\mathbf{D}^{\mathrm{A}}$ и $\mathbf{D}^{\mathrm{R}}$) один за другим «порождаются» некоторым источником данных и распределение значений очередного элемента x_n зависит только от текущего состояния источника, которое, в свою очередь, определяется значением предыдущего элемента x_{n-1} последовательности. Назовем элемент последовательности, «порожденный» источником в некотором состоянии, элементом этого состояния. Элементы каждого состояния будем кодировать/декодировать независимо. Это возможно, поскольку в момент обработки данного элемента значение предыдущего элемента известно не только кодеру, но и декодеру (декодирование осуществляется без задержки). Эффективность кодирования зависит от выбора множества состояний.

Определим способ построения состояний в стиле [5]. Выберем некоторое натуральное число t — порог. Отнесем к нулевому состоянию те элементы x_n, для которых $n \geq 1$ и $|x_{n-1}| < t$, прочие элементы отнесем к первому состоянию. Теперь последовательность $\mathbf{x}$ разложена на две подпоследовательности элементов нулевого и первого состояний $\mathbf{x}_0$ и $\mathbf{x}_1$. Квазиэнтропия и скорость кодирования пары подпоследовательностей равны

$$H(\{\mathbf{x}_0,\mathbf{x}_1\}) = \frac{N_0}{N}H(\mathbf{x}_0) + \frac{N_1}{N}H(\mathbf{x}_1);$$
$$R(\{\mathbf{x}_0,\mathbf{x}_1\}) = \frac{N_0}{N}R(\mathbf{x}_0) + \frac{N_1}{N}R(\mathbf{x}_1), \quad (8)$$

где N_0 и N_1 — число элементов нулевого и первого состояний, а квазиэнтропия и скорость кодирования каждой отдельной подпоследовательности $\mathbf{x}_0$ и $\mathbf{x}_1$ даются формулами (4) и (3).

Результирующие квазиэнтропия и скорость кодирования (8) зависят от выбора порога t, определяющего состояния. При любом значении порога квазиэнтропия (8) не превышает квазиэнтропии всей последовательности $\mathbf{x}$ (4) (см., например, [4]). Естественная оптимизационная задача — нахождение порога, при котором квазиэнтропия (8) принимает минимальное значение, — может быть решена для конкретных данных численно путем перебора.

Применим описанный метод разложения данных на два состояния к ошибкам предсказания $\mathbf{D}^{\mathrm{A}}$ и $\mathbf{D}^{\mathrm{R}}$. Оптимальные пороги в обоих случаях принимают одинаковые значения, равные 3, 4, 4, 3, 4 для массивов I,. . . ,V. Значения оптимальной квазиэнтропии $H[2,\mathrm{opt}] \doteq H(\{\{\mathbf{D}^{\mathrm{A}}{}_0,\mathbf{D}^{\mathrm{A}}{}_1,\},\{\mathbf{D}^{\mathrm{R}}{}_0,\mathbf{D}^{\mathrm{R}}{}_1,\}\})$ для массивов I–V представлены в строке 2 табл. 1. Среднее по массивам уменьшение по сравнению с квазиэнтропией $H[1]$ составляет 0,0866 бт/п, минимальное — 0,0215 бт/п (массив V), максимальное — 0,1930 бт/п (массив III).

Нахождение оптимальных порогов для данного массива силовых кривых требует значительного времени счета и не может быть реализовано на этапе кодирования в режиме реального времени. Поэтому для построения состояний в алгоритме, предназначенном для практического применения, следует использовать общие фиксированные значения порогов для всей совокупности подлежащих сжатию массивов. В строке 3 табл. 1 представлены значения квазиэнтропии $H[2,\mathrm{fix}]$, отвечающие построенным с порогами $t = 4$ двум состояниям ошибок предсказания $\mathbf{D}^{\mathrm{A}}$ и $\mathbf{D}^{\mathrm{R}}$, для массивов I–V. Увеличение квазиэнтропии по сравнению с оптимальным значением $H[2,\mathrm{opt}]$ составляет 0,0025 бт/п для массива I, пренебрежимо мало для массива IV и равно нулю для остальных массивов. Таким образом, использование общих фиксированных порогов для построения состояний не приводит к значительному увеличению квазиэнтропии по сравнению с оптимальными значениями.

Скорости кодирования $R[2,\mathrm{fix}|0] \doteq R(\{\{\mathbf{D}^{\mathrm{A}}{}_0,\mathbf{D}^{\mathrm{A}}{}_1,\},\{\mathbf{D}^{\mathrm{R}}{}_0,\mathbf{D}^{\mathrm{R}}{}_1,\}\})$ массивов I–V алгоритмом A[2,fix|0], который независимо кодирует элементы состояний ошибок предсказания $\mathbf{D}^{\mathrm{A}}$

и $\mathbf{D}^{\mathrm{R}}$, построенных с фиксированными порогами, принимающими значение $t = 4$, и использует формулу (2) с весом $w = 2$ для построения кодовых распределений, представлены в строке 7 табл. 1. Для всех массивов имеет место снижение скорости кодирования по сравнению с алгоритмом A[1|0], которое лишь немного меньше, чем уменьшение квазиэнтропии H[2,fix] по сравнению с квазиэнтропией H[1]. Среднее по массивам снижение составляет 0,0855 бт/п, минимальное — 0,0207 бт/п (массив V), максимальное — 0,1921 бт/п (массив III). Избыточность кодирования алгоритма составляет 0,0012–0,0026 бт/п и в 1,2–1,8 раза больше избыточности алгоритма A[1|0].

Рассмотрим метод уменьшения избыточности кодирования, основанный на выборе веса w в формуле (2) для кодовых распределений. Отметим, что снижение скорости кодирования при таком подходе заведомо не превысит избыточности кодирования с принятым по умолчанию значением веса $w = 2$. Оптимизационная задача нахождения веса w, при котором для конкретной последовательности данных минимальна скорость кодирования (3), может быть решена численно путем перебора. В колонках табл. 2, обозначенных I,. . . ,V, представлены значения оптимальных весов w для ошибок предсказания $\mathbf{D}^{\mathrm{A}}$ и $\mathbf{D}^{\mathrm{R}}$ каждого из массивов I–V и элементов состояний $\mathbf{D}^{\mathrm{A}}{}_0$, $\mathbf{D}^{\mathrm{A}}{}_1$, $\mathbf{D}^{\mathrm{R}}{}_0$ и $\mathbf{D}^{\mathrm{R}}{}_1$ этих ошибок; состояния построены с фиксированными порогами, принимающими значение $t = 4$ (см. выше).

Обозначим через A[1|opt] алгоритм независимого кодирования ошибок предсказания $\mathbf{D}^{\mathrm{A}}$ и $\mathbf{D}^{\mathrm{R}}$ с использованием оптимальных весов для построения кодовых распределений. Скорости кодирования R[1|opt] массивов I–V этим алгоритмом приведены в строке 5 табл. 1. Избыточность кодирования по сравнению с алгоритмом A[1|0] уменьшается в 2,1–4,4 раза и составляет теперь 0,0003–0,0006 бт/п в зависимости от массива. Соответствующее снижение скорости кодирования составляет 0,0005–0,0017 бт/п.

Вычисление оптимальных весов требует значительного времени. Поэтому в предназначенном для применения на практике алгоритме следует использовать общие фиксированные значения весов. Такие значения для совокупности массивов I–V представлены в колонках табл. 2, обозначенных I–V.

Обозначим через A[1|fix] алгоритм независимого кодирования ошибок предсказания $\mathbf{D}^{\mathrm{A}}$ и $\mathbf{D}^{\mathrm{R}}$ с использованием фиксированных весов для построения кодовых распределений. Скорости кодирования R[1|fix] массивов I–V этим алгоритмом, приведенные в строке 6 табл. 1, практически не отличаются от скоростей кодирования R[1|opt] алгоритма A[1|opt] с оптимальными весами.

Таблица 2 Оптимальные и фиксированные веса

Массив	I	II	III	IV	V	I–V
$\mathbf{D}^{\mathrm{A}}$	30	22	14	11	13	12
$\mathbf{D}^{\mathrm{A}}{}_0$	37	13	15	15	17	15
$\mathbf{D}^{\mathrm{A}}{}_1$	22	34	32	13	16	20
$\mathbf{D}^{\mathrm{R}}$	59	285	81	81	42	45
$\mathbf{D}^{\mathrm{R}}{}_0$	29	54	22	22	46	30
$\mathbf{D}^{\mathrm{R}}{}_1$	68	332	96	80	46	60

Обозначим через A[2,fix|opt] алгоритм, который независимо кодирует элементы состояний $\mathbf{D}^{\mathrm{A}}{}_0$, $\mathbf{D}^{\mathrm{A}}{}_1$, $\mathbf{D}^{\mathrm{R}}{}_0$ и $\mathbf{D}^{\mathrm{R}}{}_1$ ошибок предсказания, построенных с фиксированными порогами, принимающими значение $t = 4$, и использует оптимальные веса для построения кодовых распределений. Скорости кодирования R[2,fix|opt] массивов I–V этим алгоритмом приведены в строке 8 табл. 1. Избыточность кодирования по сравнению с алгоритмом A[2,fix|0] уменьшается в 2,5–3,8 раза до 0,0004–0,0008 бт/п в зависимости от массива. Снижение скорости кодирования составляет 0,0009–0,0020 бт/п.

Обозначим через A[2,fix|fix] алгоритм, аналогичный A[2,fix|opt], но использующий фиксированные веса для построения кодовых распределений. Скорости кодирования R[2,fix|fix] массивов I–V этим алгоритмом, приведенные в строке 9 табл. 1, практически не отличаются от скоростей кодирования R[2,fix|opt] алгоритма A[2,fix|opt], использующего оптимальные веса.

4 Заключение

Рассмотрен ряд алгоритмов обратимого сжатия массивов силовых кривых, основанных на универсальном арифметическом кодировании ошибок предсказания, и получены оценки скорости кодирования таких алгоритмов. Показано, что разложение ошибок предсказания на независимо кодируемые вычислимые состояния и выбор подходящих весов в формуле для кодовых распределений позволяют уменьшить скорость кодирования.

Применимым на практике алгоритмом, обеспечивающим наименьшую скорость кодирования, оказался алгоритм A[2,fix|fix]. Скорости кодирования этим алгоритмом массивов I–V даны в строке 9 табл. 1. Алгоритм A[2,fix|fix] дает заметно меньшую скорость кодирования по сравнению с алгоритм обратимого сжатия стандарта JPEG 2000 [1]. Для массивов I,. . . ,V выигрыш составляет 0,1737, 0,1967, 0,1591, 0,1404, 0,1117 бт/п соответственно.

Интересно применить использованные методы универсального кодирования (разделение на вычислимые состояния, выбор весов в формуле (2)) в случае арифметического кодирования компонент вейвлет-преобразования [1]. Это должно стать предметом следующей работы.

Литература

1. *Стефанович А. И., Сушко Д. В.* О сжатии данных массивов силовых кривых // Информационные процессы, 2020. Т. 20. № 3. С. 284–296.
2. *Butt H.-J., Cappella B., Kappl M.* Force measurements with the atomic force microscope: Technique, interpretation and applications // Surf. Sci. Rep., 2005. Vol. 59. P. 1–152. doi: 10.1016/j.surfrep.2005.08.003.
3. *Witten I. H., Neal R., M., Cleary J. G.* Arithmetic coding for data compression // Commun. ACM, 1987. Vol. 30. No. 6. P. 520–540. doi: 10.1145/214762.214771.
4. *Штарьков Ю. М.* Универсальное кодирование. Теория и алгоритмы. — М.: Физматлит, 2013. 288 с.
5. *Сушко Д. В., Штарьков Ю. М.* О сжатии томографических данных // Информационные процессы, 2008. Т. 8. № 4. С. 240–255.

Поступила в редакцию 30.12.2020

COMPRESSION ALGORITHMS FOR FORCE VOLUME DATA I: CODING OF PREDICTION ERRORS

D. V. Sushko

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The author considers the problem of reversible (lossless) compression of force volume data which are the three-dimensional arrays with 16-bit integer elements. Such arrays are the result of atomic force microscopy scanning of microobjects in the force mapping mode. The author proposes reversible compression algorithms of force volume data based on the universal arithmetic coding of their prediction errors. The author uses two methods of universal coding. The first method based on the statistical model of the source with the calculable sequence of states implies the decomposition of an error prediction sequence into two subsequences which are coded independently. The second method implies a choice of the appropriate weight while constructing the code probabilities used in arithmetic coding. The author constructs bit rate estimations for the proposed algorithms for five test arrays. The results show that combination of the universal coding methods mentioned above makes significant reduction of the bit rate. The bit rates of the most efficient algorithm among proposed practically applicable algorithms for the test arrays are 3.9285, 3.5268, 3.5024, 4.2813, and 4.2246 bit/pixel.

Keywords: atomic force microscope; force volume data; reversible compression; arithmetic coding; universal coding

DOI: 10.14357/19922264210212

References

1. Stefanovich, A. I., and D. V. Sushko. 2020. O szhatii dannykh massivov silovykh krivykh [On data compression of force volumes]. *Informatsionnye protsessy* [Information Processes] 20(3):284–296.
2. Butt, H.-J., B. Cappella, and M. Kappl. 2005. Force measurements with the atomic force microscope: Technique, interpretation and applications. *Surf. Sci. Rep.* 59:1–152. doi: 10.1016/j.surfrep.2005.08.003.
3. Witten, I. H., R. M. Neal, and J. G. Cleary. 1987. Arithmetic coding for data compression. *Commun. ACM* 30(6):520–540. doi: 10.1145/214762.214771.
4. Shtar'kov, Yu. M. 2013. *Universal'noe kodirovanie. Teoriya i algoritmy* [Universal coding. Theory and algorithms]. Moscow: Fizmatlit. 288 p.
5. Sushko, D. V., and Yu. M. Shtar'kov. 2008. O szhatii tomograficheskikh dannykh [On tomography data compression]. *Informatsionnye Protsessy* [Information Processes] 8(4):240–255.

Received December 30, 2020

Contributor

Sushko Dmitry V. (b. 1962) — Candidate of Science (PhD) in physics and mathematics, senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; dsushko@ipiran.ru

ПРИНЦИПЫ СТРУКТУРИЗАЦИИ СТАТЕЙ В ЭЛЕКТРОННЫХ СЛОВАРЯХ*

А. А. Гончаров[1], И. М. Зацман[2]

Аннотация: Рассмотрены две задачи, возникающие при переводе бумажных словарей в электронную форму представления: (1) структуризация унаследованных и существующих в бумажной форме словарных статей, обеспечивающая расширение функциональных возможностей электронного словаря по сравнению с бумажным; (2) замена традиционных способов шрифтового выделения структурных элементов словарной статьи на способы, обеспечивающие их программную адресацию в базе данных. Показано, что структуру словарных статей, используемую в традиционной лексикографии, необходимо детализировать и одновременно с этим категоризировать часть структурных элементов для расширения функциональных возможностей электронного словаря. Описан подход к формированию классификационной системы, интегрированной в электронный словарь, и последующей рубрикации структурных элементов словарных статей на ее основе. Предлагаемые решения позволяют значительно расширить функционал электронного словаря по сравнению с его бумажным аналогом и преодолеть ограничения традиционной лексикографии, обусловленные бумажной формой представления.

Ключевые слова: принципы структуризации; электронный словарь; электронная лексикография; классификационная система

DOI: 10.14357/19922264210213

1 Введение

Еще в 2000 г. французский лингвист Б. Серкилини (B. Cerquiglini) в ходе своего выступления на Седьмой международной конференции «Journée des dictionnaires», тема которой звучала как «От бумажных словарей к словарям электронным», выделил в развитии электронной лексикографии три этапа. Первый из них — создание бумажных словарей с использованием компьютера; второй — перевод существующих бумажных словарей в электронный формат; третий (на тот момент только начинавшийся) — изначальная разработка словарей в электронном формате с пользовательскими функциями, реализация которых невозможна при издании словарей на бумажном носителе [1, с. 188], например использование потоковых объектов мультимедиа.

В том же 2000 г. Р. Вешлер и К. Питтс сравнили электронные и бумажные словари применительно к обучению английскому языку как иностранному. Вывод, который они сделали, оказался неутешительным: «электронные словари по-прежнему остаются по своей сути бумажными словарями, записанными на электронный носитель» [2]. В 2012 г., по мнению С. Гранже, это утверждение — несмотря на улучшение ситуации — все еще оставалось актуальным для значительного числа электронных словарей [3, с. 2].

Сегодня также приходится признавать, что работа по созданию электронных словарей, которые принципиально расширяли бы функционал словарей бумажных и были бы удобны для пользователя и лексикографа, не теряет актуальности. Более того, при создании электронного словаря на основе бумажного необходимо учитывать отличия как в структуре их словарных статей, так и в способах выделения структурных элементов. Например, в бумажных словарях для этой цели может использоваться шрифтовое выделение (курсив, полужирный шрифт, заглавные буквы), символы шрифта Wingdings (🗐, 🏳, 🕮 и т. д.), нумерация арабскими и/или римскими цифрами и спецсимволы (точка с запятой, пробелы, скобки, слеши, кавычки и т. п.) [4].

Следовательно, когда при подготовке электронного словаря используются наследуемые лексикографические ресурсы, возникают две задачи: (1) более детальная (по сравнению с бумажным словарем) структуризация словарных статей, обеспечивающая расширение его функциональных возможностей; (2) замена традиционных способов выделения структурных элементов (= полей) раз-

*Работа выполнена в Институте проблем информатики ФИЦ ИУ РАН при поддержке РФФИ (проект 20-012-00166).

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, a.gonch48@gmail.com

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, izatsman@yandex.ru

Пример распределения содержания словарной статьи из [5] по зонам

Заглавное слово (= лемма) и его варианты			**können**
Зона грамматической информации о лемме в целом			*vmod* (*perf* hat können, *в неполных предложениях, где пропущен инфинитив полнозначного глагола* hat gekonnt)
Зона значения	Значение 1	Толкование леммы в данном значении	*для выражения потенциальной возможности*
		Варианты перевода леммы в данном значении	мочь, иметь возможность; можно; *под отрицанием* нельзя
		Примеры употребления леммы в данном значении	ich habe heute frei und kann dich besuchen я сегодня свободен и могу к тебе зайти; [. . .]
		Устойчивые конструкции с использованием леммы в данном значении*	. . ., **ich kann dir sagen!** (*только в постпозиции и с прямым порядком слов*) . . ., просто фантастика!; das war eine Schlägerei, ich kann dir sagen ну и драка была, я тебе скажу!; [. . .]
	. . .	. . .	. . .
Зона идиоматики			[. . .] **(gut) können** (*mit jmdm.*) *разг.* быть в (дружеских) отношениях (*с кем-л.*); die beiden können einfach nicht miteinander отношения у них просто не складываются; [. . .]

* В более ранних работах, в том числе [9, 10], эта зона носила название «Грамматическая фразеология с использованием леммы в данном значении».

меткой тегами и/или использование баз данных в процессе его подготовки. Лишь после их решения появляется возможность существенного расширения функциональных возможностей электронного словаря по сравнению с бумажным.

Цель статьи состоит в описании принципов решения первой задачи на примере двуязычного (немецко-русского) словаря, создаваемого группой лексикографов под руководством Д. О. Добровольского [5], с применением надкорпусной базы данных (НБД) немецких модальных глаголов, созданной в ФИЦ ИУ РАН (см. подробнее [6]). Решение этой задачи с применением НБД необходимо, в частности, и для того, чтобы обеспечить возможность фиксации ретроспективы изменений, вносимых в словарные статьи лексикографами (см. об этом [7, 8][1]). Частично рассматривается решение и второй задачи, которая более детально описана в работе [4][2].

2 Структуризация словарных статей

В работе [9, с. 92] в форме таблицы была представлена структура статьи, используемая в бумажном словаре [5][3]. Каждая строка таблицы соответствовала зоне словарной статьи с точки зрения традиционной лексикографии, а столбцы показывали уровень вложенности зон.

В таблице проиллюстрировано, каким образом содержание словарной статьи распределено по зонам. Для примера взята статья на один из модальных глаголов немецкого языка — глагол *können*. В целях экономии места в таблице приведено только первое из девяти его значений, а также сокращено число примеров употребления, устойчивых конструкций и идиом (пропуски отмечены как [. . .]).

Хотя на первый взгляд может показаться, что такого распределения содержания статьи по зонам — структурным элементам верхнего уровня — достаточно для работы со словарем в электронном формате, почти каждая зона с содержательной точки зрения может быть разделена на поля — структурные элементы нижних уровней, что даст возможность расширить спектр областей поиска в электронном словаре.

Так, «зона грамматической информации о лемме в целом» объединяет как минимум два поля: (1) «*vmod*» (‘модальный глагол’) — информация о том, к какой части речи принадлежит лемма; (2) «*perf* hat können, *в неполных предложениях, где пропущен инфинитив полнозначного глагола*

[1] В этих работах механизм фиксации изменений, вносимых в словарные статьи, рассматривается на примере лишь одной зоны статьи — зоны значения, приводимой, кроме того, в сокращенном виде. Однако при условии, что словарные статьи были структурированы, применение НБД позволяет фиксировать ретроспективу изменений, вносимых в любую из зон статьи.

[2] В этой работе задача решается на примере немецко-русского фразеологического словаря.

[3] Более подробное описание см. в [10, с. 40–44]; хотя там говорится о структуре статьи нового большого немецко-русского словаря, она во многом совпадает со структурой статьи словаря [5]. Другие зоны, которые также используются при составлении словарей, перечисляются в [11].

hat gekonnt» — информация об особенностях образования грамматических форм леммы. Более того, внутри второго поля можно выделить элемент «*perf*» (‘перфект’), указывающий, формы какого грамматического времени глагола образуются не по общему правилу.

Зона вариантов перевода леммы в данном значении в примере из таблицы распадается на 3 поля, в бумажном оформлении разделенные точкой с запятой. Таким образом сгруппированы наиболее близкие по значению варианты перевода: (1) «мочь, иметь возможность»; (2) «можно»; (3) «*под отрицанием* нельзя». Кроме того, третье поле содержит комментарий, в данном случае — объяснение условий, при которых следует использовать этот вариант перевода («*под отрицанием*»).

Зона примеров употребления леммы в данном значении, во-первых, состоит из отдельных примеров (в бумажном оформлении также разделенных точкой с запятой), в каждом из которых, во-вторых, можно выделить оригинальный текст примера и его перевод (отделенные друг от друга пробелом).

Зона «Устойчивые конструкции с использованием леммы в данном значении» имеет сходную структуру, которая, однако, включает и другие поля. Полужирным шрифтом выделена сама устойчивая конструкция, для которой в скобках курсивом могут указываться (1) синтаксические валентности[1] и (2) комментарии особенностей употребления. Более того, для некоторых устойчивых конструкций приводятся примеры их употребления (оригинал и перевод).

Зона идиоматики с точки зрения структуры почти полностью повторяет зону «Устойчивые конструкции. . .», однако идиомы могут сопровождаться указаниями на стилистические особенности их употребления — стилистическими пометами (см. «*разг.*» в нижней строке таблицы).

Из сказанного выше можно сделать вывод, что для создания электронного словаря, функционал которого был бы шире функционала соответствующего бумажного словаря, требуется сделать пригодными для адресации и программной обработки не только традиционно выделяемые зоны статьи, но и вложенные в них поля. Так, если представить пример из таблицы в виде XML-дерева[2] с учетом предлагаемого уровня детализации, получим следующий результат структуризации словарной статьи[3]:

```
⟨entry⟩
⟨hdw⟩können⟨/hdw⟩
⟨grinf⟩
  ⟨pos⟩vmod⟨/pos⟩
  ⟨grforms⟩
    ⟨form⟩perf ⟨/form⟩
      hat können, в неполных предложениях, где
      пропущен инфинитив полнозначного глагола
      hat gekonnt
  ⟨/grforms⟩
⟨/grinf⟩
⟨mnfld⟩
  ⟨mn⟩
    ⟨mndesc⟩
      ⟨interp⟩для выражения потенциальной
      возможности⟨/interp⟩
      ⟨trnsl⟩
        ⟨tgroup⟩мочь, иметь возможность⟨/tgroup⟩
        ⟨tgroup⟩можно⟨/tgroup⟩
        ⟨tgroup⟩
          ⟨comm⟩под отрицанием⟨/comm⟩ нельзя
        ⟨/tgroup⟩
      ⟨/trnsl⟩
    ⟨/mndesc⟩
    ⟨exfld⟩
      ⟨ex⟩
        ⟨orig⟩ich habe heute frei und kann dich
        besuchen⟨/orig⟩
        ⟨trnsl⟩я сегодня свободен и могу к тебе
        зайти⟨/trnsl⟩
      ⟨/ex⟩
      . . .
    ⟨/exfld⟩
    ⟨phrasfld⟩
      ⟨phras⟩
        ⟨orig⟩
        . . ., ich kann dir sagen!
          ⟨comm⟩только в постпозиции и с прямым
          порядком слов⟨/comm⟩
        ⟨/orig⟩
        ⟨trnsl⟩. . ., просто фантастика!⟨/trnsl⟩
        ⟨exfld⟩. . .⟨/exfld⟩
      ⟨/phras⟩
      . . .
```

[1] Синтаксической валентностью называется «способность слова вступать в синтаксические связи с другими элементами» [12, с. 79–80]. В нижней строке таблицы для идиомы «(gut) können» указана валентность «*mit jmdm.*» — буквально ‘с кем-либо’. Валентность «*с кем-л.*» указана и для перевода данной идиомы на русский язык.

[2] Идея использования XML-разметки словарных статей не нова и ранее была описана в контексте обмена словарными ресурсами [13, 14, с. 291–329].

[3] Разметка выполнена для статьи из таблицы с использованием следующих тегов: ⟨entry⟩ — словарная статья; ⟨hdw⟩ — лемма; ⟨grinf⟩ — грамматическая информация; ⟨pos⟩ — информация о том, к какой части речи принадлежит лемма; ⟨grforms⟩ — описание особенностей образования грамматических форм; ⟨form⟩ — грамматическая форма; ⟨usn⟩ — стилистические пометы; ⟨idmfld⟩ — зона идиоматики; ⟨idm⟩ — идиома; ⟨mnfld⟩ — зона значения; ⟨mn⟩ — значение; ⟨mndesc⟩ — описание значения; ⟨interp⟩ — толкование значения; ⟨orig⟩ — текст оригинала; ⟨trnsl⟩ — текст перевода; ⟨tgroup⟩ — группа вариантов перевода; ⟨phrasfld⟩ — зона «Устойчивые конструкции. . .»; ⟨phras⟩ — устойчивая конструкция; ⟨comm⟩ — комментарий; ⟨exfld⟩ — зона примеров; ⟨ex⟩ — пример; ⟨val⟩ — зона синтаксических валентностей.

```
      ⟨/phrasfld⟩
    ⟨/mn⟩
    ...
  ⟨/mnfld⟩
  ⟨idmfld⟩
    ⟨idm⟩
      ⟨orig⟩(gut) können
        ⟨val⟩mit jmdm.⟨/val⟩
        ⟨usn⟩разг.⟨/usn⟩
      ⟨/orig⟩
      ⟨trnsl⟩
        быть в (дружеских) отношениях
        ⟨val⟩с кем-л.⟨/val⟩
      ⟨/trnsl⟩
      ⟨exfld⟩...⟨/exfld⟩
    ⟨/idm⟩
    ...
  ⟨/idmfld⟩
  ⟨/entry⟩
```

Такая структура является более детальной, чем представленное в таблице традиционное деление на зоны, что и обеспечивает существенное расширение функционала электронного словаря. Наиболее очевидная новая возможность — поиск словарных статей по тексту любого из структурно выделенных полей.

Другие возможные объекты поиска: все статьи, где описываются единицы, которые можно перевести на русский словом «можно»; все статьи, которые включают интересующий пользователя или лексикографа структурный элемент (например, зону «Устойчивые конструкции...» или зону идиоматики) и т. д. Это может быть ценным для лексикографии (создание словарей разных типов), для обучения иностранному языку (отбор материала по значениям грамматических признаков), а также для решения переводческих задач.

3 Формирование классификационной системы

Хотя одно только выделение новых структурных элементов статьи расширяет функционал электронного словаря по сравнению с бумажным, выполнение категоризации этих элементов способно обеспечить решение еще более широкого круга задач. Категоризация структурного элемента (= поля) словарной статьи — это отнесение его к некоторому классу или группе согласно некоторому признаку, например: (1) «часть речи» (значения признака: n — существительное, v — глагол и т. д.); (2) «постоянные грамматические характеристики» (отметим, что наборы значений этого признака отличаются в зависимости от значения признака «часть речи», см. об этом также [13, с. 116]: так, для существительного это грамматический род — мужской (m), средний (n) или женский (f); для глагола — переходность (vt) или непереходность (vi) и т. п.); (3) «стилистические особенности употребления» (значения признака: *разг.*, *груб.* и т. д.) и др.

Значения признаков могут быть объединены в фасеты, которые, в свою очередь, объединяются в фасетную классификацию (с помощью которой выполняется рубрикация всей словарной статьи и/или ее структурных элементов).

Создание и использование такой классификации даст возможность искать по значению признака, например, словарные статьи, где лемма представляет собой: слово разговорного стиля; существительное среднего рода; прилагательное, имеющее особенности образования форм сравнительной степени, и т. д. Важно отметить, что могут отбираться статьи, где, например, к разговорному стилю относится не лемма, а только идиома с этой леммой (см. идиому с пометой «*разг.*» в нижней строке таблицы).

Существует также возможность добавления новых классификационных признаков. Одним из таких признаков, имеющих особую ценность для пользователей, может стать признак «семантика леммы». Для его использования следует, во-первых, выбрать готовую (или создать новую) классификационную систему, которая будет использоваться в электронном словаре, и, во-вторых, добавить поле для значения признака «семантика леммы», указывающего на принадлежность леммы к некоторому семантическому классу[1]. Таким образом, двуязычный словарь приобретает отдельные свойства словаря идеографического или тезауруса [16, 17].

Элементы реализации идеи включения в словарь семантической классификационной системы можно найти во французско-русской лексикографии — это так называемые «комплексные словарные статьи» в [18]. В отличие от традиционных статей двуязычного словаря комплексные статьи решают специализированные задачи: в них могут разъясняться трудности перевода применительно к описываемой паре языков (в случае словаря [18] это пара «французский–русский»), рассматриваться способы выражения семантических категорий (например, «цель», «причина»). Также в словарь могут включаться семантические группы слов (названия и перевод месяцев, дней недели, стран света и т. п.). Однако в рамках бумажного словаря лексикографы сталкиваются с бóльшими сложностями при реализации этой идеи, чем при создании словаря электронного.

[1] Применительно к фразеологическому словарю об этом говорилось в [15].

Включение в электронный словарь фасетной классификации с возможностью отбирать словарные статьи по значениям разных признаков и их сочетаниям дает пользователю широкие возможности поиска и существенно увеличивает спектр решаемых лексикографических задач[1].

4 Заключение

Существующие средства информатики дают возможность значительно расширить функционал электронных словарей по сравнению с бумажными словарями, записанными на электронный носитель. Однако для использования накопленных лексикографических ресурсов требуется выполнить структуризацию наследуемых словарных статей, обеспечивающую последующее наполнение баз данных наследуемыми лексикографическими ресурсами, формирование электронных словарей и выполнение в них лексико-грамматических видов поиска.

Для создания лексикографических баз знаний с развитыми возможностями семантического поиска необходимо предварительно сформировать и потом использовать лингвистическую фасетную классификацию, объединяющую грамматические, функционально-стилистические и семантические признаки с их простановкой как в словарных статьях, так и в их структурных элементах. В настоящее время проблема создания лексикографических баз знаний с подобными возможностями находится на начальной стадии решения.

Литература

1. *Pruvost J.* Des dictionnaires papier aux dictionnaires électroniques: VIIe Journée des dictionnaires (22 mars 2000): Rapport de colloque // Int. J. Lexicogr., 2000. Vol. 13. Iss. 3. P. 187–193. doi: 10.1093/ijl/13.3.187.
2. *Weschler R., Pitts Chr.* An experiment using electronic dictionaries with EFL students. http://iteslj.org/Articles/Weschler-ElectroDict.html.
3. Electronic lexicography / Eds. S. Granger, M. Paquot. — Oxford University Press, 2012. 517 p.
4. *Вакуленко В. В., Зацман И. М.* Наследуемые лексикографические ресурсы базы данных фразеологического словаря // Системы и средства информатики, 2021. Т. 31. № 2. С. 129–138.
5. Немецко-русский словарь актуальной лексики / Под ред. Д. О. Добровольского. — М.: Лексрус, 2021 (в печати).
6. *Добровольский Д. О., Зализняк Анна А.* Немецкие конструкции с модальными глаголами и их русские соответствия: проект надкорпусной базы данных // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам Междунар. конф. «Диалог». — М.: РГГУ, 2018. Вып. 17(24). С. 172–184.
7. *Гончаров А. А., Зацман И. М., Кружков М. Г.* Эволюция классификаций в надкорпусных базах данных // Информатика и её применения, 2020. Т. 14. Вып. 4. С. 108–116.
8. *Гончаров А. А., Зацман И. М., Кружков М. Г.* Представление новых лексикографических знаний в динамических классификационных системах // Информатика и её применения, 2021. Т. 15. Вып. 1. С. 82–89.
9. *Гончаров А. А., Зацман И. М., Кружков М. Г.* Темпоральные данные в лексикографических базах знаний // Информатика и её применения, 2019. Т. 13. Вып. 4. С. 90–96.
10. *Добровольский Д. О.* Беседы о немецком слове. — М.: Языки славянской культуры, 2013. 744 с.
11. *Lehmann Chr.* Lexicography. Microstructure: Structure of a lexical entry. https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/index.html.
12. Языкознание: Большой энциклопедический словарь / Гл. ред. В. Н. Ярцева. — 2-е изд. — М.: Большая Российская энциклопедия, 1998. 685 с.
13. *Ide N., Kilgarriff A., Romary L.* A formal model of dictionary structure and content // 9th EURALEX Congress (International) Proceedings. — Stuttgart: Institut für Maschinelle Sprachverarbeitung, 2000. P. 113–126.
14. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.2.1. — TEI Consortium, 2021. https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.
15. *Вакуленко В. В., Гончаров А. А., Дурново А. А., Зацман И. М.* Задачи базы данных фразеологического словаря и стадии ее проектирования // Системы и средства информатики, 2020. Т. 30. № 2. С. 113–123.
16. WordNet: An electronic lexical database / Ed. Chr. Fellbaum. — Cambridge, MA, USA: MIT Press, 1998. 423 p.
17. *Лукашевич Н. В.* Тезаурусы в задачах информационного поиска. — М.: Изд-во Московского ун-та, 2011. 512 с.
18. *Гак В. Г., Триомф Ж.* Французско-русский словарь активного типа. — М.: Русский язык, 1991. 1056 с.
19. *Успенский В. А.* Невто́н–Ньюто́н–Ны́тон, или Сколько сторон имеет языковой знак? // Сб. к 60-летию Андрея Анатольевича Зализняка «Русистика. Славистика. Индоевропеистика». — М.: Индрик, 1996. С. 598–659.

Поступила в редакцию 14.04.2021

[1] Возможность отбирать словарные статьи в зависимости от семантики описываемых в них единиц позволит проверить, описаны ли эти единицы аналогичным образом и не пропущена ли какая-либо из единиц. Такая ситуация была детально рассмотрена В. А. Успенским на примере включения в толковые словари русского языка названий букв русского алфавита в работе [19, с. 605–609].

STRUCTURING PRINCIPLES OF ELECTRONIC DICTIONARY'S ENTRIES

A. A. Goncharov and I. M. Zatsman

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Two tasks that arise when converting paper dictionaries into an electronic form are considered. In the first place, the authors suggest structuring inherited dictionary entries which provides the enrichment of the electronic dictionary's functionality, and in the second place, replacing the decorative design of the structural elements of dictionary entries with tagging that provide their addressing in databases. It is shown that the structure of dictionary entries used in traditional lexicography should be detailed. Simultaneously, it is necessary to categorize some of the structural elements to enrich the electronic dictionary's functionality. An approach to creating a classification system integrated into an electronic dictionary and classifying dictionary entries' structural items is described. The proposed solutions allow to significantly enrich the electronic dictionary's functionality compared to its paper version and overcome traditional lexicography limitations related to the paper form of dictionary representation.

Keywords: structuring principles; electronic dictionary; electronic lexicography; classification system

DOI: 10.14357/19922264210213

Acknowledgments

The study was conducted at the Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences with financial support from the Russian Foundation for Basic Research (grant No. 20-012-00166).

References

1. Pruvost, J. 2000. Des dictionnaires papier aux dictionnaires électroniques. VIIe Journée des dictionnaires (22 mars 2000). Rapport de colloque. *Int. J. Lexicogr.* 13(3):187–193. doi: 10.1093/ijl/13.3.187.
2. Weschler, R., and Chr. Pitts. An experiment using electronic dictionaries with EFL students. Available at: http://iteslj.org/Articles/Weschler-ElectroDict.html (accessed May 17, 2021).
3. Granger, S., and M. Paquot, eds. 2012. *Electronic lexicography*. Oxford University Press. 517 p.
4. Vakulenko, V. V., and I. M. Zatsman. 2021. Nasleduemye leksikograficheskie resursy bazy dannykh frazeologicheskogo slovarya [Inheritable lexicographic resources of the phraseological dictionary database]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 31(2):129–138.
5. Dobrovol'skiy, D. O., ed. 2021 (in press). *Nemetsko-russkiy slovar' aktual'noy leksiki* [German–Russian dictionary of actual vocabulary]. Moscow: Leksrus.
6. Dobrovol'skiy, D. O., and A. A. Zaliznyak. 2018. Nemetskie konstruktsii s modal'nymi glagolami i ikh russkie sootvetstviya: proekt nadkorpusnoy bazy dannykh [German constructions with modal verbs and their Russian correlates: A supracorpora database project]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po mat-lam Mezhdunar. konf. "Dialog"* [Computational Linguistics and Intellectual Technologies. Papers from the Annual Conference (International) "Dialogue"]. Moscow. 17(24):172–184.
7. Goncharov, A. A., I. M. Zatsman, and M. G. Kruzhkov. 2020. Evolyutsiya klassifikatsiy v nadkorpusnykh bazakh dannykh [Evolution of classifications in supracorpora databases]. *Informatika i ee Primeneniya — Inform. Appl.* 14(4):108–116.
8. Goncharov, A. A., I. M. Zatsman, and M. G. Kruzhkov. 2021. Predstavlenie novykh leksikograficheskikh znaniy v dinamicheskikh klassifikatsionnykh sistemakh [Representation of new lexicographical knowledge in dynamic classification systems]. *Informatika i ee Primeneniya — Inform. Appl.* 15(1):82–89.
9. Goncharov, A. A., I. M. Zatsman, and M. G. Kruzhkov. 2019. Temporal'nye dannye v leksikograficheskikh bazakh znaniy [Temporal data in lexicographic databases]. *Informatika i ee Primeneniya — Inform. Appl.* 13(4):90–96.
10. Dobrovol'skiy, D. O. 2013. *Besedy o nemetskom slove* [Studies on German lexis]. Moscow: Yazyki slavyanskoy kul'tury. 744 p
11. Lehmann, Chr. Lexicography. Microstructure: Structure of a lexical entry. Available at: https://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/index.html (accessed May 17, 2021).
12. Yartseva, V. N., ed. 1998. *Yazykoznanie: Bol'shoy entsiklopedicheskiy slovar'* [Linguistics. Great encyclopedic dictionary]. 2nd ed. Moscow: Bol'shaya Rossiyskaya entsiklopediya. 685 p.

13. Ide, N., A. Kilgarriff, and L. Romary. 2000. A formal model of dictionary structure and content. *9th EURALEX Congress (International) Proceedings*. Stuttgart: Institut für Maschinelle Sprachverarbeitung. 113–126.
14. TEI P5: Guidelines for electronic text encoding and interchange. Version 4.2.1. Available at: https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf (accessed May 17, 2021).
15. Vakulenko, V. V., A. A. Goncharov, A. A. Durnovo, and I. M. Zatsman. 2020. Zadachi bazy dannykh frazeologicheskogo slovarya i stadii ee proektirovaniya [Tasks of the phraseological dictionary database and stages of its design]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 30(2):113–123.
16. Fellbaum, Ch. 1998. *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press. 423 p.
17. Loukachevitch, N. V. 2011. *Tezaurusy v zadachakh informatsionnogo poiska* [Thesauri in information retrieval tasks]. Moscow: Izd-vo Moskovskogo un-ta. 512 p.
18. Gak, V. G., and Zh. Triomf. 1991. *Frantsuzsko-russkiy slovar' aktivnogo tipa* [French–Russian dictionary of the active type]. Moscow: Russkiy yazyk. 1056 p.
19. Uspenskiy, V. A. 1996. Nevtón–N'yutón–N'yúton, ili Skol'ko storon imeet yazykovoy znak? [Nevtón–N'yutón–N'yúton, or How many sides does a linguistic sign have?]. *Sbornik k 60-letiyu Andreya Anatol'evicha Zaliznyaka "Rusistika. Slavistika. Indoevropeistika"* [A collection of writings in honour of the 60th birthday of Andrey A. Zaliznyak "Russian studies. Slavic studies. Indo-European studies"]. Moscow: Indrik. 598–659.

Received April 14, 2021

Contributors

Goncharov Alexander A. (b. 1994) — junior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; a.gonch48@gmail.com

Zatsman Igor M. (b. 1952) — Doctor of Science in technology, Head of Department, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; izatsman@yandex.ru

ИЗВЛЕЧЕНИЕ ЗНАНИЙ О СРЕДСТВАХ ВЫРАЖЕНИЯ ЛОГИКО-СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ПРИ ПОМОЩИ НАДКОРПУСНОЙ БАЗЫ ДАННЫХ

А. А. Гончаров[1], О. Ю. Инькова[2]

Аннотация: Цель статьи — показать продуктивность использования параллельных текстов и их аннотирования в надкорпусной базе данных (НБД) коннекторов для извлечения знаний об альтернативных средствах выражения логико-семантических отношений (ЛСО). На примере наиболее известных дискурсивно аннотированных корпусов — Penn Discourse Treebank (PDTB), Prague Dependency Treebank (PDT) и Rhetorical Structure Theory Discourse Treebank (RST-DT) — авторы показывают, что в существующих исследованиях нет консенсуса относительно того, какие языковые средства относить к классу коннекторов (прототипических показателей ЛСО), а какие — к альтернативным средствам. В исследовании продемонстрировано, что применение сопоставительного метода и использование возможностей НБД коннекторов позволяет не только извлекать новое знание о средствах выражения ЛСО в изучаемых языках, но и создавать тезаурусы таких средств, в том числе альтернативных коннекторам. Кроме того, информация, хранящаяся в НБД, дает возможность получать новые знания о том, какие ЛСО могут быть выражены неспециализированными средствами и какова частотность использования этих средств для каждого ЛСО в каждом из изучаемых языков.

Ключевые слова: надкорпусная база данных; логико-семантические отношения; коннекторы; извлечение новых знаний; параллельные тексты

DOI: 10.14357/19922264210214

1 Вводные замечания

Логико-семантические, или, шире, дискурсивные, отношения, обеспечивающие связность текста на естественном языке, привлекают внимание лингвистов и специалистов по информатике уже не один десяток лет: первые исследования начали появляться в 1970-х гг. (например, работы Дж. Хоббса [1, 2]). Однако многие вопросы до сих пор остаются дискуссионными: это, в первую очередь, и само понятие «дискурсивное отношение», и понятие «коннектор» (единицы этого класса считаются прототипическими эксплицитными показателями таких отношений). Нет консенсуса и относительно того, можно ли создать исчерпывающий список коннекторов для исследуемого языка. Тем не менее важность списков коннекторов для разработки дискурсивных парсеров и, шире, средств автоматической обработки текста и автоматического извлечения информации из текста подчеркивается в ряде работ (см., например, [3, с. 55]). В [4], где описываются результаты разработки дискурсивного парсера для русского языка, отмечается, что наличие показателя ЛСО служит наиболее надежным признаком для определения того, каким именно отношением связаны фрагменты текста.

В то же время в [5] показано, что в зависимости от типа текста и вида ЛСО коннекторы используются лишь в 30%–40% случаев. Остальные случаи представляют собой либо имплицитные ЛСО (подробнее об этом понятии см., например, [6]), либо ЛСО, показателем которого являются языковые средства, отличные от коннекторов. Следовательно, качество результатов автоматической обработки текстов на естественном языке непосредственно зависит от уровня наших знаний не только о коннекторах, но и об этих, альтернативных, средствах выражения ЛСО. Цель статьи состоит в том, чтобы показать продуктивность использования параллельных текстов и поисковых возможностей НБД коннекторов, разработанной в ИПИ ФИЦ ИУ РАН (подробнее см. [7–9]), для извлечения знаний об альтернативных средствах выражения ЛСО.

2 Существующие подходы

Отправной точкой для активных исследований средств выражения ЛСО, альтернативных коннек-

[1] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, a.gonch48@gmail.com

[2] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, olyainkova@yandex.ru

торам, стала статья [10], отражающая подход создателей Пенсильванского дискурсивно аннотированного корпуса (PDTB) [11] к этому вопросу. При аннотировании корпуса выяснилось, что если два фрагмента текста связаны каким-либо ЛСО (дискурсивным отношением в терминах PDTB), то это отношение может: (*i*) выражаться коннектором[1]; (*ii*) не выражаться коннектором, причем какой-либо коннектор можно добавить (= имплицитные ЛСО); (*iii*) не выражаться коннектором, причем никакой коннектор не может быть добавлен из-за возникающей в этом случае семантической избыточности. Авторы пришли к выводу, что такая избыточность вызвана наличием альтернативных коннекторам лексических средств выражения ЛСО — «альтернативных лексикализаций» (alternative lexicalizations, AltLex).

Если в PDTB в разряд коннекторов попадает ограниченный круг языковых единиц, то разработчики Пражского корпуса синтаксических зависимостей (PDT) трактуют понятие коннектор более широко, включая в этот класс большинство лексических средств, которые так или иначе могут выражать ЛСО [3, 12]. Коннекторы при этом разделяются на «первичные» (primary) и «вторичные» (secondary), довольно разнообразные по своей морфологической природе. Не включаются в число коннекторов лишь так называемые «неуниверсальные», или «свободные связующие сочетания» (non-universal / free connecting phrases), образующие третий класс средств выражения ЛСО. Ср. (1)–(4) из [3, с. 51, 68]:

(1) Fred didn't stop joking. **As a result**, his friends enjoyed hilarity throughout the evening.

'Фред не переставал шутить. **В результате** его друзья смеялись весь вечер'.[2]

(2) I had all the necessary qualifications. **Despite this**, I didn't get the job.

'Я удовлетворял всем квалификационным требованиям. **Несмотря на это**, на работу меня не приняли'.

(3) Fred didn't stop joking. This **caused** hilarity among his friends for the whole evening.

'Фред не переставал шутить. Это **вызывало** смех его друзей весь вечер'.

(4) Fred has pneumonia. **Because of this illness**, he will be absent from his work for two weeks.

'У Фреда пневмония. **Вследствие этой болезни** его не будет на работе две недели'.

Так, в (1) причинно-следственные ЛСО выражены союзом *as a result* — «первичным коннектором». В (2) и (3) используются «вторичные коннекторы»: в (2) это сочетание предлога *despite* с анафорическим выражением *this*, отсылающим к ситуации *I had all the necessary qualifications*, которое выражает уступительные ЛСО; в (3) «вторичным коннектором» считается глагол *caused*, выражающий причинные ЛСО. Наконец, в (4) *because of this illness* рассматривается как «неуниверсальное», или «свободное связующее сочетание», поскольку оно непосредственно связано с предыдущим контекстом (в котором упомянута пневмония Фреда), в отличие от *despite this* в (3), имеющего более общее значение.

«Альтернативным лексикализациям» в подходе PDTB соответствуют второй и третий классы единиц в PDT [3, с. 54]. Такого же широкого подхода к «коннекторам» придерживаются разработчики русского дискурсивно аннотированного корпуса [13].

Подчеркнем, что во всех описанных случаях во внимание принимаются лишь лексические средства выражения ЛСО. В [3, с. 62] даже особо отмечается, что «коннекторами» не считаются синтаксические и морфологические средства, например относительные придаточные или деепричастия, которые в ряде языков способны выражать ЛСО (см. ниже). Подчеркивая важность создания лексиконов связующих средств, разработчики Пражского корпуса не дают, тем не менее, списка «вторичных коннекторов».

В последней версии PDTB (3.0) появился класс показателей ЛСО AltLexC (где «C» означает *Construction*), включающий лексико-синтаксические средства выражения ЛСО [14, с. 9–10, 76]. Однако если для «первичных коннекторов» приводится список языковых единиц, то ни для «альтернативных лексикализаций», ни для нового класса AltLexC списков не дается. В [14, с. 75–76] приводятся лишь ЛСО, которые выражают те единицы этих классов, что зафиксированы в последней версии корпуса PDTB.

В рамках Теории риторической структуры (Rhetorical Structure Theory, RST) принципы классификации языковых единиц, способных выражать риторические (в данной терминологии) отношения, даны в работе [15, с. 8, 9], которая служит пособием по аннотированию показателей риторических отношений в корпусе RST Discourse Treebank (RST-DT). Поскольку объем понятия «риторическое отношение» шире, чем ЛСО, так как включает не только отношения связности, которые могут быть выражены коннектором, то и набор показателей этих отношений шире. К «первичным» коннекторам добавляются показатели самой разнообразной природы: лексические, морфологические

[1] К коннекторам относятся сочинительные и подчинительные союзы, а также некоторые другие языковые единицы, за которыми грамматиками английского языка традиционно признается связующая функция.

[2] Здесь и далее в отсутствие других указаний перевод авторов статьи.

(временны́е формы), семантические (синонимия, антонимия и др.), синтаксические (различные виды придаточных и др.), графические (знаки препинания и др.) и т. д.; причем эти средства могут также комбинироваться друг с другом[1]. Однако в пособии по аннотированию приводятся лишь типы показателей риторических отношений, а не их список.

3 Альтернативные средства выражения логико-семантических отношений в надкорпусной базе данных коннекторов

Если во всех упомянутых выше работах средства выражения ЛСО изучаются на одноязычном материале, то НБД коннекторов позволяет проводить исследования на материале параллельных текстов, используя методы сопоставительной лингвистики. Аннотирование употреблений коннекторов в параллельных текстах позволило заметить, что в некоторых случаях использованный в оригинале коннектор переведен не коннектором, а другим языковым средством (или наоборот, в оригинале коннектор отсутствует, но появляется в переводе). С точки зрения сопоставительного подхода такие случаи представляют собой примеры «дивергентного перевода»[2].

Для обозначения языковых единиц, не являющихся коннекторами, но способными выражать ЛСО, предлагается использовать термин «альтернативные коннекторам средства выражения ЛСО». В НБД такие средства делятся на (*i*) лексические; (*ii*) грамматические и (*iii*) пунктуационные.

3.1 Лексические средства

В примере (5) коннектор *то есть*, выражающий ЛСО переформулирования, двумя переводчиками передан альтернативными коннекторам лексическими средствами — ‘я имею в виду’ и ‘я хочу сказать’ соответственно.

(5) Моя теща, **то есть** мать жены моей, тоже ничего не видит. [Н. В. Гоголь. Нос (1832–1833)]

‘Ma belle-mère, **j’entends** la mère de ma femme, a, elle aussi, la vue faible.’ [Tr. H. Mongault (1938)]

‘Ma belle-mère, **je veux dire**, la mère de ma femme, elle non plus, elle n’y voit rien du tout.’ [Tr. A. Markowitz (2007)]

3.2 Грамматические средства

В примере (6) для передачи коннектора *потому что*, выражающего ЛСО причины, также два переводчика используют форму причастия настоящего времени, которая во французском языке способна выражать это отношение.

(6) В заключение прибавлял, что он «был бы счастлив, если б удалось ему на себе оправдать свое убеждение, но что достичь этого он не надеется, **потому что** это очень трудно». [И. А. Гончаров. Обломов (1848–1859)]

‘Et il concluait en ajoutant qu’il serait tout à fait heureux s’il parvenait à justifier ses idées par son comportement, mais qu’il n’espérait pas y parvenir, cette adéquation **étant** fort difficile à atteindre.’ [Tr. A. Adamov (1959)]

‘En guise de conclusion il ajoutait «qu’il serait heureux s’il pouvait justifier ses convictions par sa propre vie, mais qu’il ne l’espérait pas, cet objectif **étant** trop difficile à atteindre».’ [Tr. L. Jurgenson (1988)]

3.3 Пунктуационные средства

В (7) уже упоминавшийся выше коннектор *потому что* передан в двух переводах двоеточием.

(7) . . .но коллежский асессор Ковалев не мог слышать запаха, **потому что** закрылся платком и потому что самый нос его находился бог знает в каких местах. [Н. В. Гоголь. Нос (1832–1833)]

‘. . .Mais l’assesseur de collège Kovaliov ne pouvait pas s’en rendre compte**:** il avait caché son visage sous un mouchoir, et d’ailleurs son nez se trouvait en cet instant Dieu sait où.’ [Tr. B. de Schloezer (1925)]

‘. . .Cependant le major Kovaliov ne s’en trouvait point incommodé**:** il tenait son mouchoir sur son visage, et d’ailleurs son nez se promenait. . . Dieu sait où.’ [Tr. H. Mongault (1938)]

Тот факт, что в примерах (5)–(7) несколько переводчиков выбирают альтернативные средства выражения ЛСО, свидетельствует о том, что эти средства выражают ЛСО на регулярной основе, а не являются единичными переводческими решениями.

Промежуточное положение между коннекторами и альтернативными им средствами выражения ЛСО занимают языковые единицы, представляющие собой сочетание коннектора и лексического и/или грамматического средства. Так, в (8) при

[1] Заметим, что в более ранних версиях корпусов, размеченных в соответствии с RST, во всех таких случаях риторическое отношение считалось имплицитным. Согласно версии, представленной в [15], оно оказывается эксплицитным, а в тех случаях, когда никакой из потенциальных показателей отношения не может быть идентифицирован, проставляется метка *unsure*.

[2] Термин «дивергентный перевод» заимствован из работы [16], посвященной использованию многоязычных корпусов в контрастивных исследованиях; он был уточнен для исследования коннекторов в [17].

переводе коннектора *так как* на французский язык использовано сочетание коннектора *puisque* и формы причастия настоящего времени. Для обозначения таких сочетаний используется термин «комбинированные средства выражения ЛСО», а переводное соответствие считается конгруэнтно-дивергентным.

(8) Предприятия, расположенные в городе и севернее города, не выполнили своих обязательств перед государством, **так как** находятся в районе военных действий. [В. С. Гроссман. Жизнь и судьба (1960)]

'Les entreprises situées dans la ville, ou un peu au nord, n'avaient pu remplir leurs obligations envers l'État, **puisque se trouvant** en pleine zone d'opérations militaires.' [Tr. A. Berelowitch (1980)]

Похожая группа появляется в последней версии PDTB (3.0) — «AltLex Relations Linked with Explicits» [14, с. 80]. Она, однако, не аналогична классу комбинированных средств выражения ЛСО, так как включает единицы, относимые нами к коннекторам, такие как *and in general, but in general*.

В табл. 1 приводятся данные (по состоянию на 02.03.2021) о числе зафиксированных в НБД коннекторов переводных соответствий (далее — ПС), где в русском языке (языке оригинала) исследуемая языковая единица является коннектором. Для сравнения: в Пражском корпусе на «вторичные коннекторы», при довольно широкой трактовке этого термина, приходится 5% [12, с. 456], а в PDTB 3.0 на AltLex и AltLexC — в сумме чуть более 3% [14, с. 5].

По данным НБД, в дивергентных ПС коннектор чаще всего передается лексическими средствами, а реже всего — знаками препинания. На грамматические средства, которые совсем не учитываются в Пражском корпусе и лишь недавно стали аннотироваться в PDTB, приходится более 26% альтернативных средств. Этим, видимо, можно объяснить более высокую долю дивергентных соответствий в НБД. В табл. 2 приводятся наиболее употребительные грамматические средства выражения ЛСО. На данный момент они не разделены на более мелкие подклассы, но фасетная классификация, используемая в НБД (см. [18]), позволяет решить эту задачу.

В табл. 3 сравниваются, с одной стороны, данные о том, какие ЛСО могут передаваться с использованием альтернативных средств при переводе с русского языка на французский, и, с другой стороны, данные о том, какие ЛСО могут выражаться альтернативными средствами в англоязычном корпусе PDTB 3.0. В работе [14] отсутствуют данные об общем числе примеров для каждого ЛСО, поэтому в табл. 3 приводятся только абсолютные цифры. В четвертом столбце табл. 3 указано общее число примеров для AltLex и AltLexC, так как обе группы соответствуют альтернативным средствам выражения ЛСО в НБД. Данные из табл. 3 показывают, что использование параллельных текстов позволяет извлечь знания об альтернативных средствах выражения большего числа ЛСО, чем анализ одноязычного материала.

Таблица 1 Виды ПС для коннекторов русского языка в НБД (направление перевода русский–французский)

Всего	Конгруэнтное ПС	Дивергентное ПС	Конгруэнтно-дивергентное ПС	Эксплицитная языковая единица отсутствует
11 175 (100%)	8 948 (80,07%)	942 (8,43%)	167 (1,5%)	1 118 (10%)

Таблица 2 Наиболее употребительные дивергентные ПС, зафиксированные в НБД (направление перевода русский–французский)

№	Средство выражения ЛСО в переводе	Число ПС (с коннектором в оригинале)
1	Придаточное определительное предложение	47
2	Форма деепричастия настоящего времени	35
3	Конструкция с местоименным повтором	20
4	Форма деепричастия настоящего времени в сочетании с *tout*	20
5	Форма причастия настоящего времени	17
. . .	. . .	. . .
46	*Il n'y a que. . . qui*	1
		246

Таблица 3 Альтернативные коннекторам средства выражения в НБД и в PDTB

Отношение в НБД	**Дивергентных ПС**	**Отношение в PDTB**	**AltLex**
Исключение Исключение из рассмотрения	131 27	Exception.Arg2-as-excpt	3
«Вопреки ожидаемому» Сопоставительные «Вопреки ожидаемому» иллокутивные Возместительное противопоставление Противительно-уступительные иллокутивные Противительно-уступительные Контраст Противопоставление	47 35 19 12 4 5 2 2	Contrast	41
Аддитивные иллокутивные Аддитивные пропозициональные Соединительные	45 16 28	Conjunction	139
Замещение	73	Substitution.Arg1-as-subst; Substitution.Arg2-as-subst	29
Переформулирование	68	Equivalence	10
Спецификация	60	Instantiation.Arg2-as-instance; Level-of-detail.Arg2-as-detail	106
Временные	53	Asynchronous.Precedence; Asynchronous.Succession; Synchronous	160
Временные метаязыковые	4		
Пропозициональное сопутствование	42	—	0
Уступительные	42	Concession.Arg1-as-denier; Concession.Arg2-as-denier	39
Условные Метаязыковые условные	36 1	Condition.Arg1-as-cond; Condition.Arg2-as-cond	74
Коррекция	32	—	0
Пропозициональная причина	32	Cause.Reason	281
Сравнительные	31	Similarity	63
Неединственности	25	—	0
Аналогия	18	—	0
Иллокутивная причина	16	Cause+Belief.Reason+Belief	6
Иллокутивное сопутствование	16	—	0
Пропозициональная альтернатива Гипотетическая альтернатива	15 1	Disjunction	0
Экстенсиональная генерализация Интенсиональная генерализация	7 5	Instantiation.Arg1-as-instance	1
Тождество	11	—	0
Несоответствие	6	—	0
Обобщающее переформулирование	6	Level-of-detail.Arg1-as-detail	16
Отрицательная альтернатива Оговорка	4 2	Negative-condition.Arg2-as-negCond	2
Следствие	4	Cause.Result; Cause.negResult; Cause+Belief.Result+Belief	663
Уступительные иллокутивные	4	Concession+SpeechAct.Arg2-as-denier+SpeechAct	1
Отрицание тождества	1	—	0
Цель	0	Purpose.Arg1-as-goal; Purpose.Arg2-as-goal	35
—	0	Manner.Arg1-as-manner; Manner.Arg2-as-manner	3
	988		1672

4 Заключительные замечания

Таким образом, исследование ЛСО с использованием параллельных текстов и аннотирование их показателей в НБД позволяет, во-первых, извлекать новое знание о средствах выражения ЛСО (в том числе альтернативных коннекторам) и создавать тезаурусы таких средств в изучаемых языках; во-вторых, на основе информации, хранящейся в НБД, получать новые знания о том, какие ЛСО могут быть выражены неспециализированными средствами и какова частность их использования для каждого ЛСО в каждом из изучаемых языков. Все это способно улучшить работу дискурсивных парсеров за счет пополнения спектра признаков, на основании которых принимаются решения о наличии того или иного ЛСО между фрагментами текста.

Литература

1. *Hobbs J. R.* A computational approach to discourse analyses. — New York, NY, USA: Department of Computer Science, City College, City University of New York, 1976. Research Report 76-2.
2. *Hobbs J. R.* Why is discourse coherent? — Menlo Park, CA, USA: SRI International, 1978. SRI Technical Note 176.
3. *Danlos L., Rysová K., Rysová M., Stede M.* Primary and secondary discourse connectives: Definitions and lexicons // Dialogue Discourse, 2018. Vol. 9. No. 1. P. 50–78.
4. *Chistova E. V., Shelmanov A. O., Kobozeva M. V., Pisarevskaya D. B., Smirnov I. V., Toldova S. Yu.* Classification models for RST discourse parsing of texts in Russian // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам ежегодной Междунар. конф. «Диалог». — М.: РГГУ, 2019. Вып. 18(25). С. 163–176.
5. *Taboada M.* Discourse markers as signals (or not) of rhetorical relations // J. Pragmatics, 2006. Vol. 38. No. 4. P. 567–592.
6. *Гончаров А. А., Инькова О. Ю.* Имплицитные логико-семантические отношения и метод их поиска в параллельных текстах // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам ежегодной Междунар. конф. «Диалог». — М.: РГГУ, 2020. Вып. 19(26). С. 310–320.
7. *Зацман И. М., Инькова О. Ю., Кружков М. Г., Попкова Н. А.* Представление кроссязыковых знаний о коннекторах в надкорпусных базах данных // Информатика и её применения, 2016. Т. 10. Вып. 1. С. 106–118.
8. *Зацман И., Кружков М., Лощилова Е.* Методы и средства информатики для описания структуры неоднословных коннекторов // Структура коннекторов и методы ее описания / Под ред. О. Ю. Иньковой. — М.: ТОРУС ПРЕСС, 2019. С. 205–230.
9. Семантика коннекторов: количественные методы описания / Под ред. О. Иньковой. — Bern/Berlin: Peter Lang, 2021. 276 с.
10. *Prasad R., Joshi A., Webber B.* Realization of discourse relations by other means: Alternative lexicalizations // 23rd Conference (International) on Computational Linguistics Proceedings. — Beijing, China, 2010. P. 1023–1031. https://www.aclweb.org/anthology/C10-2118.pdf.
11. Penn Discourse Treebank Project (PDTB). https://www.seas.upenn.edu/~pdtb.
12. *Rysová M., Rysová K.* The centre and periphery of discourse connectives // 28th Pacific Asia Conference on Language, Information and Computing Proceedings. — Phuket: Department of Linguistics, Chulalongkorn University, 2014. P. 452–459. https://www.aclweb.org/anthology/Y14-1052.pdf.
13. Ru-RSTreebank. Русскоязычный дискурсивный корпус. https://rstreebank.ru.
14. *Webber B., Prasad R., Lee A., Joshi A.* The Penn Discourse Treebank 3.0: Annotation Manual, 2019. https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf.
15. *Das D., Taboada M.* RST Signalling Corpus: Annotation Manual, 2014. https://www.sfu.ca/~mtaboada/docs/publications/RST_Signalling_Corpus_Annotation_Manual.pdf.
16. *Johansson S.* Seeing through multilingual corpora: On the use of corpora in contrastive studies. — Amsterdam/Philadelphia: John Benjamins, 2007. 377 p.
17. *Инькова О. Ю.* Аннотирование параллельных текстов: понятие «дивергентный перевод» // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам ежегодной Междунар. конф. «Диалог». — М.: РГГУ, 2019. Вып. 18(25). С. 227–238.
18. *Зацман И. М., Инькова О. Ю., Нуриев В. А.* Построение классификационных схем: методы и технологии экспертного формирования // Научно-техническая информация. Сер. 2: Информационные процессы и системы, 2017. № 1. С. 8–22.

Поступила в редакцию 06.04.2021

EXTRACTING KNOWLEDGE ABOUT MEANS OF EXPRESSION OF LOGICAL-SEMANTIC RELATIONS FROM THE SUPRACORPORA DATABASE

A. A. Goncharov and O. Yu. Inkova

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The goal of this paper is to demonstrate how parallel texts annotated with a supracorpora database (SCDB) can be efficiently used to extract knowledge about alternative means of expression of logical-semantic relations (LSR). The authors review the most prominent discursively annotated corpora (Penn Discourse Treebank, Prague Dependency Treebank, and Rhetorical Structure Theory Discourse Treebank) to support the observation that there is no consensus among the researchers as to which linguistic means are to be considered connectives (i. e., prototypical markers of LSR) and which means are deemed "alternative." The research shows that application of the comparative method while leveraging the capabilities of the SCDB of connectives makes it possible not only to extract new knowledge about LSR markers but also to create thesauri of various means of LSR expression in the languages involved, including the alternative ones. In addition, the SCDB data makes it possible to generate new knowledge on correlations between specific LSRs and unconventional means of LSR expression and calculate frequencies of utilization of these means for the studied languages.

Keywords: supracorpora database; logical-semantic relations; connectives; knowledge generation; parallel texts

DOI: 10.14357/19922264210214

References

1. Hobbs, J. R. 1976. A computational approach to discourse analyses. New York, NY: Department of Computer Science, City College, City University of New York. Research Report 76-2.
2. Hobbs, J. R. 1978. Why is discourse coherent? Menlo Park, CA: SRI International. SRI Technical Note 176.
3. Danlos, L., K. Rysová, M. Rysová, and M. Stede. 2018. Primary and secondary discourse connectives: Definitions and lexicons. *Dialogue Discourse* 9(1):50–78.
4. Chistova, E. V., A. O. Shelmanov, M. V. Kobozeva, D. B. Pisarevskaya, I. V. Smirnov, and S. Yu. Toldova. 2019. Classification models for RST discourse parsing of texts in Russian. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po mat-lam Mezhdunar. konf. "Dialog"* [Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference (International) "Dialogue"]. Moscow: RSHI. 18(25):163–176.
5. Taboada, M. 2006. Discourse markers as signals (or not) of rhetorical relations. *J. Pragmatics* 38(4):567–592.
6. Goncharov, A. A., and O. Yu. Inkova. 2020. Implitsitnye logiko-semanticheskie otnosheniya i metod ikh poiska v parallel'nykh tekstakh [Implicit logical-semantic relations and a method of their identification in parallel texts]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po mat-lam Mezhdunar. konf. "Dialog"* [Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference (International) "Dialogue"]. Moscow: RSHI. 19(26):310–320.
7. Zatsman, I. M., O. Yu. Inkova, M. G. Kruzhkov, and N. A. Popkova. 2016. Predstavlenie kross-yazykovykh znaniy o konnektorakh v nadkorpusnykh bazakh dannykh [Representation of cross-lingual knowledge about connectors in suprocorpora databases]. *Informatika i ee Primeneniya — Inform. Appl.* 10(1):106–118.
8. Zatsman, I., M. Kruzhkov, and E. Loshchilova. 2019. Metody i sredstva informatiki dlya opisaniya struktury neodnoslovnykh konnektorov [Methods and means of informatics for multiword connectives structure description]. *Struktura konnektorov i metody ee opisaniya* [Connectives structure and methods of its description]. Ed. O. Yu. Inkova. Moscow: TORUS PRESS. 205–230.
9. Inkova, O., ed. 2021. *Semantika konnektorov: kolichestvennye metody opisaniya* [Semantics of connectives: Quantitative methods of analysis]. Bern/Berlin: Peter Lang. 276 p.
10. Prasad, R., A. Joshi, and B. Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. *23rd Conference (International) on Computational Linguistics Proceedings*. Beijing, China. 1023–1031. Available at: https://www.aclweb.org/anthology/C10-2118.pdf (accessed June 15, 2021).
11. Penn Discourse Treebank Project. Available at: https://www.seas.upenn.edu/~pdtb/ (accessed May 19, 2021).
12. Rysová, M., and K. Rysová. 2014. The centre and periphery of discourse connectives. *28th Pacific Asia Conference on Language, Information and Computing Proceedings*. Phuket: Department of Linguistics, Chulalongkorn University. 452–459. Available at: https://www.aclweb.org/anthology/Y14-1052.pdf (accessed June 15, 2021).
13. Ru-RSTreebank: Russkoyazychnyy diskursivnyy korpus [Ru-RSTreebank: Russian discourse corpus]. Available at: https://rstreebank.ru/ (accessed May 19, 2021)

14. Webber, B., R. Prasad, A. Lee, and A. Joshi. 2019. The Penn Discourse Treebank 3.0: Annotation manual. Available at: https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf (accessed May 19, 2021)
15. Das, D., and M. Taboada. 2014. RST signalling corpus: Annotation manual. Available at: https://www.sfu.ca/~mtaboada/docs/publications/RST_Signalling_Corpus_Annotation_Manual.pdf (accessed May 19, 2021)
16. Johansson, S. 2007. *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Amsterdam/Philadelphia: John Benjamins. 377 p.
17. Inkova, O. Yu. 2019. Annotirovanie parallel'nykh tekstov: ponyatie "divergentnyy perevod" [Annotation of parallel texts: The concept of divergent translation]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po mat-lam Mezhdunar. konf. "Dialog"* [Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference (International) "Dialogue"]. Moscow: RSHI. 18(25):227–238.
18. Zatsman, I., O. Inkova, and V. Nuriev. 2017. The construction of classification schemes: Methods and technologies of expert formation. *Autom. Doc. Math. Linguist.* 51(1):27–41.

Received April 6, 2021

Contributors

Goncharov Alexander A. (b. 1994) — junior scientist, Institute of Informatics Problems, Federal Research Center"Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; a.gonch48@gmail.com

Inkova Olga Yu. (b. 1965) — Doctor of Science in philology, senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; olyainkova@yandex.ru

МЕТОДЫ ОЦЕНКИ КАЧЕСТВА МАШИННОГО ПЕРЕВОДА: СОВРЕМЕННОЕ СОСТОЯНИЕ

В. А. Нуриев[1], А. Ю. Егорова[2]

Аннотация: Представлен обзор современных методов оценки качества машинного перевода (МП). В основе этих методов лежат два подхода — автоматический и экспертный. Автоматическая оценка построена на сопоставлении с референтным (профессиональным/эталонным) переводом (РП). Экспертная (с привлечением человека-эксперта) учитывает в первую очередь функциональность: качество перевода оценивается в прагматико-функциональном аспекте, т. е. принимается во внимание, насколько полученный перевод справляется со своими задачами. В первой части статьи рассматривается ряд метрик, используемых для автоматической оценки МП, отмечаются их недостатки и описываются новые направления в их разработке. Вторая часть статьи сфокусирована на экспертной оценке МП. Здесь приведены несколько основных способов такой оценки: оценивание в соответствии с критериями точности и естественности, ранжирование переводов, прямое оценивание, оценка с учетом коэффициента редактирования перевода человеком, аннотирование перевода с применением типологии ошибок.

Ключевые слова: машинный перевод; качество перевода; оценка качества машинного перевода; автоматические метрики; прямое оценивание; типология ошибок машинного перевода

DOI: 10.14357/19922264210215

1 Введение

Проблема и, следовательно, необходимость оценки качества переводного текста возникает на регулярной основе, причем не только в профессиональном сообществе, но и в жизни обычного человека. Во многом это связано с тем, что МП стал неотъемлемой частью повседневной реальности. Происходит реструктуризация рынка переводческих услуг и, в частности, МП, а бюджет индустрии постоянно наращивает объемы (см. об этом [1, с. 260–263]). Ежедневно с помощью публично доступного веб-сервиса нейронного МП Google.Translate обрабатывается около 143 млрд слов в 100 языковых парах [2]. Человек использует МП для решения задач широкого профиля: для получения информации, связанной с конкретной и требующей незамедлительных действий проблемой (перевод технического сопровождения, инструкции к лекарствам и т. д.); для покупок на зарубежных сайтах; в целях оптимизации профессиональной деятельности переводчика с помощью внедрения в его рабочий цикл этапа, предполагающего последующую редактуру автоматически сгенерированного текста. Не все указанные задачи предполагают обязательный высокий уровень качества МП. Для осуществления ряда из них достаточно общего понимания содержания даже при наличии несущественных ошибок, нарушающих правила целевого языка. Для выполнения других задач требуется высокое качество полученного автоматическим способом перевода, что указывает на необходимость постоянно оценивать динамику этого качества, изучать и совершенствовать методы его оценки. Этим обусловлено повышенное внимание научного сообщества к данной проблеме: за последние четыре года были опубликованы несколько авторитетных монографий, фокусирующихся на оценке качества МП [3–5].

Целью статьи, таким образом, ставится обзор современных тенденций в разработке методов оценки качества МП. В основе этих методов лежат два подхода — автоматический и экспертный. Автоматическая оценка построена на сопоставлении с референтным (профессиональным/эталонным) переводом (*англ.* reference translation). Экспертная (с привлечением человека-эксперта) оценка учитывает в первую очередь функциональность: качество перевода оценивается тем выше, чем успешнее он справляется со своими задачами.

2 Автоматическая оценка качества машинного перевода

Как правило, автоматическая оценка измеряет уровень соответствия МП одному или нескольким РП. Чтобы определить уровень соответствия

[1] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, nurieff.v@gmail.com

[2] Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, ann.shurova@gmail.com

МП и РП, применяются критерии точности (доля правильно переведенного) и полноты (доля переведенных слов, совпадающих с профессиональным переводом). Ниже представлены некоторые метрики автоматической оценки качества МП.

Метрика, к которой обращаются чаще всего, — это разработанная в IBM метрика BLEU (Bilingual Evaluation Understudy) [6]. Она вошла в золотой стандарт автоматической оценки качества МП и нередко применяется в качестве эталонной. Сопоставление МП и РП проводится путем вычисления n-граммной точности (максимальная длина n-граммного блока слов равна 4). Чтобы избежать искажения в оценке, за слишком короткий перевод назначается штраф (brevity penalty — BP), n-граммная точность при этом представляет собой «отношение последовательностей из n слов, совпадающих в МП и РП, к общему числу последовательностей из n слов в МП» [7, с. 111–112]. «Оценка... вычисляется как произведение среднего геометрического из полученных модифицированных коэффициентов и штрафного коэффициента» [8, с. 86]. Полученное значение BLEU изменяется в пределах от 0 до 1. При процентном представлении значение изменяется в промежутке от 0% до 100%. Вычисляется метрика по следующей формуле:

$$\mathrm{BLEU} = \mathrm{BP}\exp\left(\sum_{n=1}^{n} w_i \log p_i\right),$$

$$\mathrm{BP} = \min\left(1, \frac{c}{r}\right),$$

«где w_i — положительные веса для каждого используемого параметра n-грамм... n — максимальная длина n-грамм, i — длина блока в пределах n-граммы, p_i — модифицированная точность n-грамм, c — длина полученного машинного перевода, r — длина наилучшего совпадающего эталонного текста» [8, с. 86]. Метрика BLEU использует статистические инструменты, не принимая во внимание лингвистические знания.

Другая наиболее востребованная метрика — METEOR (Metric for Evaluation of Translation with Explicit Ordering) — предусматривает интеграцию языковых знаний. Так, наряду с n-граммными совпадениями в МП и РП она учитывает изменения в словоформах, синонимические ряды и т. д. [9]. Поэтому для обеспечения ее функционирования необходимо привлечение баз данных, содержащих лингвистическую информацию, нужна морфологическая разметка и вычислительно затратное пословное выравнивание. Иначе говоря, требуется сложная и тонкая настройка, куда вовлечено гораздо больше параметров, чем в BLEU.

Еще одной получившей широкое распространение метрикой автоматической оценки качества МП стала TER (Translation Error Rate). Она исходит из расчета исправлений/трансформаций, необходимых для приведения МП к эталонному образцу, и вычисляется по следующей формуле:

$$\mathrm{TER} = \frac{\text{Число редактирований}}{\text{Средняя длина эталонных переводов}}.$$

При этом пунктуационные знаки принимаются за отдельные слова, а трансформациями считаются не только удаление, вставка и замена, но и перестановка — в отличие, например, от метрики WER (Word Error Rate), которая эту последнюю трансформацию не учитывает (подробнее о TER и WER см. в [8, 10]).

Наряду с рассмотренными метриками автоматической оценки качества МП также имеются: PER (Position-Independent Word Error Rate), chrF (Character F-measure), NIST (название образовано от US National Institute of Standards and Technology) и др. (о них см. [7, 8, 11, 12]).

Целесообразность использования метрик автоматической оценки качества МП постоянно ставится под вопрос. Действительно, может ли метрика с простейшим алгоритмом вычисления типа BLEU (как и другие метрики) адекватно отражать отличия между МП и РП? Основные критические замечания, высказываемые в этой связи, заключаются в следующем.

1. При вычислении не принимается во внимание, что слова несут на себе разную функциональную нагрузку и имеют неодинаковую релевантность для формирования предложения.
2. Сравнение РП и МП носит локальный характер и проводится на уровне n-граммного соответствия, при этом упускается из виду грамматическая связность в рамках всего предложения, что искажает результаты в пользу систем МП, которые лучше переводят отдельные словарные блоки, но не всегда способны грамматически правильно оформить целое предложение.
3. Вычисляемые значения не информативны: неизвестно, как интерпретировать значение BLEU, равное, например, 30,7%, так как при вычислении задействовано множество факторов — число РП, языковая пара, терминологическое наполнение текста, схема токенизации, используемая для вычленения слов в РП и МП.
4. Ненадежность алгоритма оценки. Так, недавние эксперименты показали, что BLEU оценивает выполненные человеком переводы на том же уровне, что и машинные, хотя последние имеют гораздо худшее качество. В ходе этих

экспериментов с помощью BLEU выполнялось сравнение между несколькими РП, а также между РП и МП.

Эти недостатки необходимо учитывать в разработке новых метрик автоматической оценки МП, как необходимо учитывать и то, что в идеальном случае оценка, получаемая с помощью такой метрики, должна демонстрировать явную корреляцию с оценкой человека-эксперта. Обычно эту корреляцию рассчитывают с помощью коэффициента корреляции Пирсона [10, с. 61]. Его значение варьируется от 0 до 1, и чем оно выше в указанном диапазоне, тем лучше метрика.

Автоматические метрики получили всеобщее признание в качестве эффективного способа для оценки продуктивности систем статистического МП, однако они не совсем приспособлены, чтобы сравнивать производительность систем МП разного типа между собой, и в этом отношении разработки средств автоматической оценки МП пока не достигли сколь-нибудь значимых результатов. Вместе с тем такие разработки интенсивно ведутся, и автоматические метрики постоянно совершенствуются.

Так, все большее распространение получает подход, учитывающий при сопоставлении МП и РП морфологические, синтаксические и семантические параметры. Это, например, MEANT, где сопоставляются синтаксические древовидные структуры и принимаются во внимание такие свойства, как семантические роли [13]. Или RIBES (Rank-based Intuitive Bilingual Evaluation Score) — метрика, специально разработанная для языковых пар типа японский–английский, где коренным образом различается синтаксическое устройство.

Имеются попытки применять машинное обучение — обучать метрики на данных, полученных по результатам оценки человеком-экспертом (см., например, BEER (BEtter Evaluation as Ranking) [14, 15] или BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) — одну из самых новых метрик, которая использует нейросетевую языковую модель BEURT и обучается на рейтинговых данных [16]).

Наряду с увеличением степени корреляции между оценкой человека-эксперта и автоматической оценкой МП разработчики также стремятся обеспечить большую информативность метрик (см. выше замечание о непрозрачности значения BLEU) и снижение вычислительной трудоемкости.

Важным сейчас становится создание информационных ресурсов, содержащих лингвистические знания разной направленности, предназначенные для обучения современных автоматизированных метрик. Примером таких ресурсов служат надкорпусные базы данных, разрабатываемые в отделе 54 ФИЦ ИУ РАН [17].

Подробнее о новейших разработках в области автоматизированных средств оценки качества МП см. [10, с. 59–64].

3 Экспертная оценка качества машинного перевода

Говоря об экспертной оценке качества МП с привлечением специалистов (лингвистов, переводчиков), можно выделить несколько основных способов: оценивание в соответствии с критериями точности и естественности, ранжирование переводов, прямое оценивание, оценка с учетом коэффициента редактирования перевода человеком, аннотирование перевода с применением типологии ошибок.

Понятие «правильности» перевода недоопределено и, следовательно, плохо применимо. Вот почему для оценки качества МП руководствуются критериями[1] точности (adequacy) и естественности (fluency), используя в опросе экспертов 5-балльную шкалу Ликерта [19]. Такой подход имеет свои недостатки: эксперты не всегда последовательны в своем выборе из-за неоднозначности определений в шкале оценки, к тому же одни специалисты более снисходительны при назначении оценок, чем другие.

Чтобы избежать этих трудностей, при оценке двух и более систем МП применяется ранжирование переводов относительно друг друга. Для измерения меры согласия между экспертами используют коэффициенты каппа Коэна [10, с. 48], каппа Флейса [20].

Так, в работе [21] ранжирование проводилось для оценки качества переводов, реализованных посредством системы статистического фразового МП (СФМП) и системы нейронного МП (НМП). В ходе эксперимента каждому эксперту были представлены триплеты, состоящие из предложения на исходном языке и двух его переводов (полученных с помощью СФМП и НМП). Экспертам предлагалось оценить триплет и приписать его к одному из трех классов, показывающих соотношение качества сравниваемых переводов:

$$\text{СФМП} = \text{НМП}\,;\ \text{СФМП} < \text{НМП};\ \text{СФМП} > \text{НМП}\,.$$

Полученные экспертные оценки были сопоставлены с результатами автоматических метрик (BLEU, TER, Character F-measure).

[1] Другие возможные критерии оценки описаны в [18].

В эксперименте, описанном в [22], помимо ранжирования еще задействованы постредактирование переводов, экспертная аннотация ошибок в МП, а также оценка точности/естественности. Подобно [21], для установления корреляции между экспертной и автоматической оценкой используются автоматические метрики.

Одной из последних разработок в области оценки качества МП является прямое оценивание (direct assessment) [10, с. 49–50]. Оно предполагает оценку одного предложения единовременно (в отличие от ранжирования переводов) с применением 100-балльной шкалы, которая имеет вид немаркированной прямой с бегунком. Для экспертов характерны неодинаковые ожидания в отношении качества МП: одни склонны его оценивать выше, а другие, наоборот, ниже, что может объясняться имеющимися предубеждениями о низком качестве МП. Кроме того, разными экспертами 5-балльная шкала используется неравномерно — некоторые никогда не ставят самый низкий и самый высокий баллы. 100-балльная шкала представляет собой более гибкий оценочный инструмент. Она дает возможность измерить ожидания в отношении качества МП у каждого эксперта с помощью среднего балла всех его оценок, выявляя задействованный интервал шкалы, который отражается в дисперсии оценок. Оценки разных экспертов нормируются согласно формулам в [10, с. 49–50]. Переводы, поступающие эксперту для обработки, генерируются в разных системах МП и выбираются случайным образом. После нормирования оценок, полученных от каждого из экспертов, вычисляется средний балл для переводов отдельно взятой системы МП.

Прямое оценивание было использовано в ходе краудсорсинговой кампании по оценке качества МП, организованной ACL (Association for Computational Linguistics) в 2018 г. в рамках Конференции по компьютерной лингвистике (Workshop on Machine Translation, WMT).

Оценивать качество перевода можно и с точки зрения усилий по его постредактированию. Так, при оценке МП с учетом HTER[1] (Human Translation Edit Rate — коэффициент редактирования перевода человеком) [10, с. 51–52] эксперты получают подборку переводов, выполненных разными системами МП, которые им предлагается отредактировать. Затем для каждой системы МП проводится сопоставление перевода с его отредактированной версией и подсчитывается число изменений, сделанных экспертом.

Качество МП может оцениваться и в процессе аннотирования перевода с применением типологии ошибок. Обзор классификаций представлен в работе [23].

Одной из наиболее известных является типология DQF/MQM (Dynamic Quality Framework — динамическая модель оценки качества; Multidimensional Quality Metrics — многомерные метрики качества), разработанная в TAUS[2] и DFKI[3] в 2014 г. [24]. Типология имеет 4 уровня: наиболее специфицированные типы ошибок относятся к четвертому уровню; при этом при оценке перевода можно выбирать степень спецификации, т. е. использовать от одного до четырех уровней в зависимости от задачи. Также в типологии учитываются четыре степени критичности ошибок. Подробнее о MQM-метриках см. в работе [7].

Типология DQF/MQM получила широкое распространение. Так, в [25] она применяется для проведения количественного анализа работы разных систем МП. При этом классификация претерпевает ряд изменений, обусловленных необходимостью учитывать особенности славянских языков (в данном случае хорватского).

Следует отметить, что типологии ошибок могут быть специфицированы в зависимости от цели исследования. Так, по мнению авторов статьи [26], категория «Терминология» в классификации MQM не отражает нюансы, которые могут возникать при ошибочном переводе терминов. В работе предпринята попытка провести анализ ошибок в переводе терминов, уточнить их классификацию и сопоставить на этой основе работу систем СФМП и НМП.

Представленная в [26] типология ошибок включает в себя 5 классов:

(1) «Ошибка в словопорядке» (Reorder error);

(2) «Ошибка в формообразовании» (Inflectional error);

(3) «Ошибка в части термина» (Partial error);

(4) «Лексическая ошибка» (Incorrect lexical selection);

(5) «Пропуск термина» (Term drop).

Оставшиеся виды ошибок образуют 6-й класс, который подразделяется на три подкласса:

– «Копирование исходного термина» (Source term copied);

[1] Аббревиатура совпадает с названием автоматической метрики HTER (Human-targeted Translation Error Rate) (подробнее см. [18, с. 25]).

[2] Translation Automation User Society — Пользовательское сообщество по автоматизации перевода.

[3] Deutsches Forschungszentrum für Künstliche Intelligenz — Немецкий центр исследований искусственного интеллекта.

– «Ошибка, вызванная затруднением при снятии многозначности слова на целевом языке» (Disambiguation issue in target);
– «Другие ошибки» (Other error).

Переводы исходных терминов, в которых не было допущено ошибок, объединены в отдельный класс «Правильного перевода» (Correct translation). В нем авторы исследования выделяют еще 7 подклассов, которые демонстрируют разнообразие моделей перевода и отображают степень соответствия переводного эквивалента исходному термину.

Еще одним примером типологии ошибок может послужить классификация, представленная в [27]. Она подробно описана в работе [28]. Эта типология имеет 5 укрупненных классов ошибок, которые, в свою очередь, делятся на подклассы.

В работе [29] проводится количественный анализ ошибок мультимодальных систем НМП, способных обрабатывать изображения. За основу взята классификация ошибок [27] с некоторыми уточнениями. Изменения в ней обусловлены интересом авторов исследования к тому, как мультимодальные системы НМП переводят «визуальные» термины (visual terms) — термины, обозначающие понятия, прямое соответствие которым можно найти на предъявляемом изображении, причем задействованы только укрупненные классы типологии ошибок [27] без уточнения их дальнейшего иерархического устройства.

С учетом всех преобразований модифицированная классификация [29] включает в себя следующие классы ошибок:

(1) «Пропущенные слова» (Missing words);
(2) «Неправильные слова» (Incorrect words), куда входят подклассы
   – «Неправильный перевод» (Mistranslation);
   – «Неправильная форма, лишние слова, стилистическая ошибка» (Incorrect form, extra words or style);
(3) «Другие ошибки» (Other), куда входят подклассы
   – «Словопорядок» (Word order);
   – «Неизвестные слова» (Unknown words);
   – «Пунктуационная ошибка» (Punctuation).

Также был добавлен новый, 4-й класс, получивший название «Визуальная категория» (Visual category). В него входят 4 подкласса:

– «Правильный перевод» (Correct);
– «Неправильный перевод» (Mistranslation);
– «Неправильный, но интересный перевод» (Incorrect but interesting);
– «Новый термин» (Novel).

Разнообразие способов экспертной оценки МП свидетельствует о неослабевающем интересе профессионального сообщества к этой области, а также говорит о том, что даже с учетом существенно меньшей стоимости и большей быстроты автоматической оценке не доверяют полностью и стремятся проверить ее с помощью мнения компетентного человека.

4 Заключение

В статье представлен обзор современных подходов к оценке качества МП. Выделены два основных направления: автоматизированная оценка и оценивание с привлечением человека-эксперта. С изменением парадигмы МП и внедрением нейросетей в архитектуру автоматических переводчиков изменяются и разработки в области оценки качества МП. Это затрагивает в первую очередь автоматические метрики, используемые для оценивания переводов: для обеспечения их работы пытаются применять машинное обучение. В качестве тренировочных привлекаются данные, полученные по результатам оценки человеком. Нововведения имеются и в области экспертной оценки качества МП. Одна из последних разработок здесь — прямое оценивание. Востребованным остается аннотирование МП с применением типологии ошибок. Оно стало одним из самых продуктивных способов оценивания, поскольку позволяет гибко типологизировать ошибки в соответствии с целым рядом параметров, которые легко варьировать в зависимости от конкретных характеристик текста, поступающего на вход системы МП.

Литература

1. *Larsonneur C.* Neural machine translation: From commodity to commons? // When translation goes digital: Case studies and critical reflections / Eds. R. Desjardins, C. Larsonneur, Ph. Lacour. — Cham, Switzerland: Palgrave Macmillan, 2021. P. 257–280.
2. *Davenport C.* Google Translate processes 143 billion words every day // Android Police, 2018. https://www.androidpolice.com/2018/10/09/google-translate-processes-143-billion-words-every-day.
3. Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Machine translation: Technologies and applications ser. — Cham, Switzerland: Springer International Publishing, 2018. Vol. 1. 292 p.

4. *Specia L., Scarton C., Paetzold G. H.* Quality estimation for machine translation. — Synthesis lectures on human language technologies ser. — London: Morgan & Claypool, 2018. 162 p.
5. *Bittner H.* Evaluating the evaluator: A novel perspective on translation quality assessment. — New York, NY, USA: Routledge, 2020. 282 p.
6. *Papineni K., Roukos S., Ward T., Zhu W. J.* BLEU: A method for automatic evaluation of machine translation // 40th Annual Meeting on Association for Computational Linguistics Proceedings. — Philadelphia, PA, USA: Association for Computational Linguistics, 2002. P. 311–318.
7. *Рычихин А. К.* О методах оценки качества машинного перевода // Системы и средства информатики, 2019. Т. 29. № 4. С. 106–118.
8. *Козина А. В., Черепков Е. А., Белов Ю. С.* Автоматические метрики оценки качества машинного перевода // Системный администратор, 2019. № 11. С. 84–87.
9. *Banerjee S., Lavie A.* METEOR: An automatic metric for MT evaluation with improved correlation with human judgments // Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics Proceedings. — Ann Arbor, MI, USA: Association of Computational Linguistics, 2005. P. 65–72.
10. *Koehn Ph.* Neural machine translation. — New York, NY, USA: Cambridge University Press, 2020. 394 p.
11. *Popović M.* chrF: Character n-gram F-score for automatic MT evaluation // 10th Workshop on Statistical Machine Translation Proceedings. — Lisboa, Portugal: Association for Computational Linguistics, 2015. P. 392–395.
12. *Popović M.* chrF deconstructed: β parameters and n-gram weights // 1st Conference on Machine Translation Proceedings. — Berlin, Germany: Association for Computational Linguistics, 2016. Vol. 2. P. 499–504.
13. *Chi-kiu Lo.* MEANT 2.0: Accurate semantic MT evaluation for any output language // Conference on Machine Translation Proceedings. — Copenhagen, Denmark: Association for Computational Linguistics, 2017. Vol. 2. P. 589–597.
14. *Stanojević M., Sima'an K.* BEER: BEtter evaluation as ranking // 9th Workshop on Statistical Machine Translation Proceedings. — Baltimore, MD, USA: Association for Computational Linguistics, 2014. P. 414–419.
15. *Stanojević M., Sima'an K.* Evaluating MT systems with BEER // Prague Bulletin Mathematical Linguistics, 2015. No. 104. P. 17–26.
16. *Sellam T., Das D., Parikh A. P.* BLEURT: Learning robust metrics for text generation // arXiv.org, 9 Apr 2020. arXiv:2004.04696 [cs.CL].
17. *Инькова О. Ю.* Надкорпусная база данных как инструмент изучения формальной вариативности коннекторов // Компьютерная лингвистика и интеллектуальные технологии: По мат-лам ежегодной Международ. конф. «Диалог». — М.: РГГУ, 2018. Вып. 17(24). С. 240–253.
18. *Castilho Sh., Doherty S, Gaspari F., Moorkens J.* Approaches to human and machine translation quality assessment // Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Cham, Switzerland: Springer, 2018. P. 9–38.
19. *Likert R.* A technique for the measurement of attitudes // Arch. Psychol., 1932. Vol. 140. P. 1–55
20. *Fleiss J. L.* Measuring nominal scale agreement among many raters // Psychol. Bull., 1971. Vol. 76. No. 5. P. 378–382.
21. *Shterionov D., Superbo R., Nagle P., et al.* Human versus automatic quality evaluation of NMT and PBSMT // Machine Translation, 2018. Vol. 32. P. 217–235.
22. *Castilho S., Moorkens J., Gaspari F., et al.* Evaluating MT for massive open online courses. A multifaceted comparison between PBSMT and NMT systems // Machine Translation, 2018. Vol. 32. P. 255–278.
23. *Popovic M.* Error classification and analysis for machine translation quality assessment // Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Cham, Switzerland: Springer, 2018. P. 129–158.
24. *Lommel A.* Metrics for translation quality assessment: A case for standardizing error typologies // Translation quality assessment: From principles to practice / Eds. J. Moorkens, Sh. Castilho, F. Gaspari, S. Doherty. — Cham, Switzerland: Springer, 2018. P. 109–127.
25. *Klubička F., Toral A., Sánchez-Cartagena V. M.* Quantitative fine-grained human evaluation of machine translation systems: A case study on English to Croatian // Machine Translation, 2018. Vol. 32. P. 195–215.
26. *Haque R., Hasanuzzaman M., Way A.* Analysing terminology translation errors in statistical and neural machine translation // Machine Translation, 2020. Vol. 34. P. 149–195.
27. *Vilar D., Xu J., D'Haro L., Ney H.* Error analysis of statistical machine translation output // 5th Conference (International) on Language Resources and Evaluation Proceedings. — Genoa, Italy: European Language Resources Association, 2006. P. 697–702.
28. *Гончаров А. А., Бунтман Н. В., Нуриев В. А.* Ошибки в машинном переводе: проблемы классификации // Системы и средства информатики, 2019. Т. 29. № 3. С. 92–103.
29. *Calixto I., Liu Q.* An error analysis for image-based multimodal neural machine translation // Machine Translation, 2019. Vol. 33. P. 155–177.

Поступила в редакцию 14.04.2021

METHODS OF QUALITY ESTIMATION FOR MACHINE TRANSLATION: STATE-OF-THE-ART

V. A. Nuriev and A. Yu. Egorova

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: The paper reviews the state-of-the-art methods of quality estimation for machine translation. These methods are grounded in two general approaches: automatic and manual. The automatic assessment builds on the data from comparison of the machine translation system output against the human-generated reference translation. The manual (human) evaluation primarily takes into account pragmatic and functional aspects: the translation quality is assessed bearing in mind how well the system output is suited to fulfill the translation tasks. The first part presents some automatic metrics for evaluation of machine translation quality. Also, it speaks about both shortcomings of such metrics and new trends in their development. The other part of the paper is focused on human evaluation of machine translation. It describes: (i) evaluation of adequacy and fluency; (ii) ranking of translations; (iii) direct assessment; (iv) computation of the human translation edit rate, and (v) translation annotation involving an error typology.

Keywords: machine translation; translation quality; evaluation of machine translation quality; automatic metrics; direct assessment; typology of machine translation errors

DOI: 10.14357/19922264210215

References

1. Larsonneur, C. 2021. Neural machine translation: From commodity to commons? *When translation goes digital: Case studies and critical reflections*. Eds. R. Desjardins, C. Larsonneur, and P. Lacour. Cham: Palgrave Macmillan. 257–280.
2. Davenport, C. 2018. Google Translate processes 143 billion words every day. *Android Police*. Available at: https://www.androidpolice.com/2018/10/09/google-translate-processes-143-billion-words-every-day/ (accessed May 5, 2021).
3. Moorkens, J., S. Castilho, F. Gaspari, and S. Doherty, eds. 2018. *Translation quality assessment: From principles to practice*. Machine translation: Technologies and applications ser. Cham: Springer International Publishing. Vol. 1. 299 p.
4. Specia, L., C. Scarton, and G. H. Paetzold. 2018. *Quality estimation for machine translation*. Synthesis lectures on human language technologies ser. London: Morgan & Claypool Publs. 162 p.
5. Bittner, H. 2020. *Evaluating the evaluator: A novel perspective on translation quality assessment*. New York, NY: Routledge. 282 p.
6. Papineni, K., S. Roukos, T. Ward, and W. J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. *40th Annual Meeting on Association for Computational Linguistics Proceedings*. Philadelphia, PA: Association for Computational Linguistics. 311–318.
7. Rychikhin, A. K. 2019. O metodakh otsenki kachestva mashinnogo perevoda [On methods of machine translation quality assessment]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(4):106–118.
8. Kozina, A. V., E. A. Cherepkov, and Yu. S. Belov. 2019. Avtomaticheskie metriki otsenki kachestva mashinnogo perevoda [Automatic metrics for machine translation evaluation]. *Sistemnyy administrator* [System Administrator] 11:84–87.
9. Banerjee, S., and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics Proceedings*. Ann Arbor, MI: Association of Computational Linguistics. 65–72.
10. Koehn, Ph. 2020. *Neural machine translation*. New York, NY: Cambridge University Press. 394 p.
11. Popović, M. 2015. chrF: Character n-gram F-score for automatic MT evaluation. *10th Workshop on Statistical Machine Translation Proceedings*. Lisboa, Portugal: Association for Computational Linguistics. 392–395.
12. Popović, M. 2016. chrF deconstructed: β parameters and n-gram weights. *1st Conference on Machine Translation Proceedings*. Berlin, Germany: Association for Computational Linguistics. 2:499–504.
13. Chi-kiu, Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. *Conference on Machine Translation Proceedings*. Copenhagen, Denmark: Association for Computational Linguistics. 2:589–597.
14. Stanojević, M., and K. Sima'an. 2014. BEER: BEtter evaluation as ranking. *9th Workshop on Statistical Machine Translation Proceedings*. Baltimore, MD: Association for Computational Linguistics. 414–419.
15. Stanojević, M., and K. Sima'an. 2015. Evaluating MT systems with BEER. *Prague Bulletin Mathematical Linguistics* 104:17–26.
16. Sellam, T., D. Das, and A. P. Parikh. 2020. BLEURT: Learning robust metrics for text generation. Available at:

https://arxiv.org/pdf/2004.04696.pdf (accessed May 5, 2021).

17. Inkova, O. Yu. 2018. Nadkorpusnaya baza dannykh kak instrument formal'noy variativnosti konnektorov [Supra-corpora database as an instrument of the study of the formal variability of connectives]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: po mat-lam ezhegodnoy Mezhdunar. konf. "Dialog"* [Computer Linguistic and Intellectual Technologies: Conference (International) "Dialog" Proceedings]. Moscow. 17(24):240–253.
18. Castilho, Sh., S. Doherty, F. Gaspari, and J. Moorkens. 2018. Approaches to human and machine translation quality assessment. *Translation quality assessment: From principles to practice*. Eds. J. Moorkens, Sh. Castilho, F. Gaspari, and S. Doherty. Cham: Springer. 9–38.
19. Likert, R. 1932. A technique for the measurement of attitudes. *Arch. Psychol.* 140:1–55.
20. Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76(5):378–382.
21. Shterionov, D., R. Superbo, P. Nagle, *et al.* 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation* 32:217–235.
22. Castilho, S., J. Moorkens, F. Gaspari, *et al.* 2018. Evaluating MT for massive open online courses. A multifaceted comparison between PBSMT and NMT systems. *Machine Translation* 32:255–278.
23. Popovic, M. 2018. Error classification and analysis for machine translation quality assessment. *Translation quality assessment: From principles to practice*. Eds. J. Moorkens, Sh. Castilho, F. Gaspari, and S. Doherty. Cham: Springer. 129–158.
24. Lommel, A. 2018. Metrics for translation quality assessment: A case for standardising error typologies. *Translation quality assessment: From principles to practice*. Eds. J. Moorkens, Sh. Castilho, F. Gaspari, and S. Doherty. Cham: Springer. 109–127.
25. Klubička, F., A. Toral, and V. M. Sánchez-Cartagena. 2018. Quantitative fine-grained human evaluation of machine translation systems: A case study on English to Croatian. *Machine Translation* 32:195–215.
26. Haque, R., M. Hasanuzzaman, and A. Way. 2020. Analysing terminology translation errors in statistical and neural machine translation. *Machine Translation* 34:149–195.
27. Vilar, D., J. Xu, L. D'Haro, and H. Ney. 2006. Error analysis of statistical machine translation output. *5th Conference (International) on Language Resources and Evaluation Proceedings*. 697–702.
28. Goncharov, A. A., N. V. Buntman, and V. A. Nuriev. 2019. Oshibki v mashinnom perevode: problemy klassifikatsii [Machine translation errors: Problems of classification]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 29(3):92–103.
29. Calixto, I., and Q. Liu. 2019. An error analysis for image-based multi-modal neural machine translation. *Machine Translation* 33:155–177.

Received April 14, 2021

Contributors

Nuriev Vitaly A. (b. 1980) — Candidate of Science (PhD) in philology, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; nurieff.v@gmail.com

Egorova Anna Yu. (b. 1991) — junior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; ann.shurova@gmail.com

СТОХАСТИЧЕСКАЯ ДИНАМИКА САМООРГАНИЗУЮЩИХСЯ СОЦИАЛЬНЫХ СИСТЕМ С ПАМЯТЬЮ (ЭЛЕКТОРАЛЬНЫЕ ПРОЦЕССЫ)

А. С. Сигов[1], Е. Г. Андрианова[2], Л. А. Истратов[3]

Аннотация: Обсуждаются вопросы применения методологии и подходов теоретической информатики для анализа и моделирования социальных групповых процессов. Обработка социологических данных электоральной кампании выборов президента США в 2016 г. позволила построить гистограммы плотности вероятности амплитуд отклонений предпочтений избирателей в зависимости от величины интервала времени их определения и разработать модель описания стохастических социальных процессов с учетом самоорганизации и наличия памяти, учитывающую основные характеристики наблюдаемых процессов. При создании модели рассмотрены схемы вероятностей переходов между возможными состояниями социальной системы и выведено нелинейное дифференциальное уравнение второго порядка. Сформулирована и решена граничная задача для определения функции плотности вероятности амплитуды отклонений предпочтений избирателей в зависимости от величины интервала времени ее определения. Дифференциальное уравнение модели содержит член, отвечающий за возможность самоорганизации, а также учитывает наличие памяти. Возможность возникновения осцилляций определяется начальными условиями. Разработанную модель можно использовать для анализа электоральных кампаний и принятия решений.

Ключевые слова: функция распределения амплитуд колебаний; стохастическая динамика; самоорганизация; наличие памяти; осцилляции плотности вероятности; электоральные процессы

DOI: 10.14357/19922264210216

1 Введение

Человеческий фактор в социальных системах активно влияет на происходящие явления, внося неопределенность воздействием на протекающие процессы и создавая возможности для самоорганизации систем. Возможность появления памяти о предыдущих состояниях системы и оказанных на нее воздействиях приводит к существенно нелинейной динамике процессов в социальных системах. Применение методов и средств теоретической информатики и кибернетики для моделирования динамики самоорганизующихся социальных систем с памятью позволяет получить качественно новые результаты для описания социальных систем.

2 Обзор исследований по анализу и моделированию социальных процессов

Социальные процессы характеризуются сложными механизмами протекания и стохастичностью. Различные множественные состояния зависят от влияния друг на друга участников процесса [1, 2]. Изначально теоретические подходы к описанию социальных систем имели много общего с кинетическим описанием физических систем [3, 4]. Эти модели актуальны, однако наблюдения процессов в социальных сетях показали ограничения кинетических моделей. В частности, кинетические модели фокусируются на появлении мгновенных глобальных каскадов, инициированных одиночными локальными возмущениями.

Известны примеры, когда главную роль играют пороговые механизмы развития процессов в социальных системах. Состояния узлов могут зависеть от внешних импульсов, например из средств массовой информации [5], определяющих стохастическую составляющую процессов.

Более поздние модели описания процессов в социальных сетях используют стохастические подходы, учитывающие зависимости состояний от времени [6–8]. В работе [6] рассматривается модель смешанного членства в стохастически формирующихся группах, основанная на рассмотрении попарных измерений, таких как присутствие или от-

[1]МИРЭА — Российский технологический университет (РТУ МИРЭА), sigov@mirea.ru
[2]МИРЭА — Российский технологический университет (РТУ МИРЭА), andrianova@mirea.ru
[3]МИРЭА — Российский технологический университет (РТУ МИРЭА), istratov@mirea.ru

сутствие связей между парой объектов. Данная модель позволяет при определенных допущениях отследить динамику изменения численности членов в формирующихся группах и кластеризацию членов по группам. В работах [7, 8] групповые социальные процессы рассматриваются с позиций теории перколяции, что позволяет учитывать структуру сети. В частности, было исследовано влияние плотности сети на величину порога ее перколяции (проводимости сети в целом) и динамику его достижения. Для моделирования поведения участников социальных процессов используют теорию многоагентных систем. На основании некоторых правил переходов агенты принимают определенные состояния, образуют связанную по своим свойствам группу, могут сотрудничать, чтобы решить некую задачу или достигнуть определенной цели [8], а поведение агентов может зависеть от динамически меняющихся условий [9].

В работе [10] использована теория клеточных автоматов для описания социальной системы, поведение которой зависит от свойств внешней среды и структуры поведения. Поведение социальной системы описывают четыре параметра: разнообразие; связность; взаимозависимость; адаптируемость. При увеличении взаимозависимости и адаптивности поведение системы становится более упорядоченным и целенаправленным. Считаем, что дальнейшая разработка моделей поведения участников социальных процессов является актуальной задачей. Рассмотрена электоральная кампания как стохастический динамический процесс перехода между возможными состояниями системы с течением времени. Изменения состояния могут иметь не только случайный характер, но и учитывать возможные процессы самоорганизации, наличие памяти и осцилляции.

Более подробное описание математических моделей динамики процессов в сложных социальных и экономических системах можно найти в обзоре [11] и ряде других оригинальных работ [12–22].

3 Выбор данных для анализа динамики социальных процессов и получение гистограмм статистических распределений и их моментов

Для создания и проверки модели требуется значительный объем наблюдаемых данных. Большая статистически достоверная база данных по электоральным процессам доступна на ресурсе http://www.realclearpolitics.com. Анализ динамики изменения настроений избирателей в ходе предвыборных кампаний и прогнозирование на ее основе итогов представляет огромный интерес. Поэтому для анализа и разработки модели выбраны электоральные процессы. Для обработки наблюдаемых данных (изменения процентов предпочтения избирателей в США на протяжении 500 дней, с 1 июля 2015 г. по 7 ноября 2016 г.) и определения функций плотности вероятности амплитуд колебаний предпочтений избирателей использован следующий алгоритм.

1. Выбираем все значения исследуемых предпочтений избирателей за некоторый диапазон времени (сутки, неделя, месяц и т. д.), вычисляем значения амплитуды изменения величины колебаний предпочтений избирателей за различные интервалы времени.

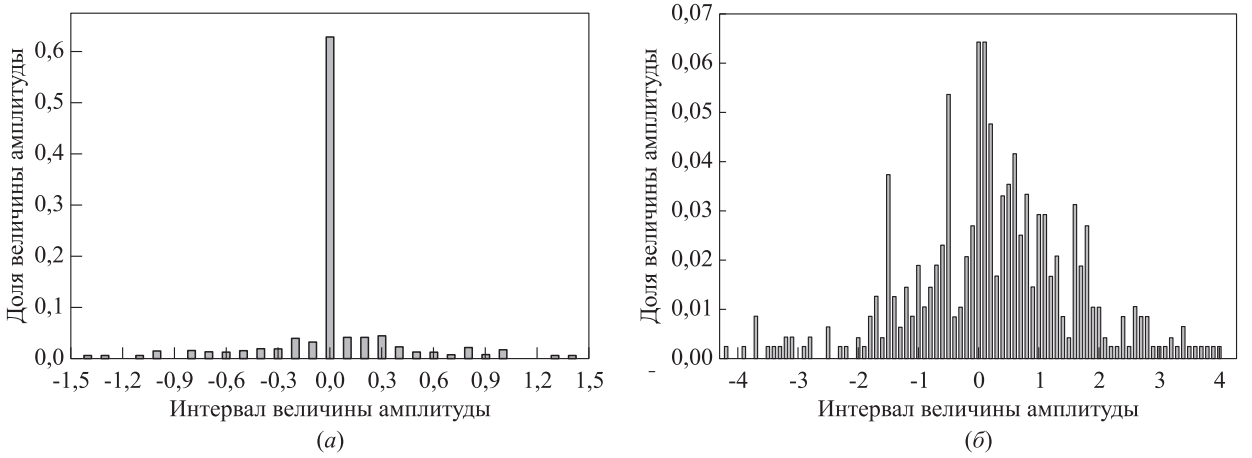


Рис. 1 Гистограммы плотности вероятности амплитуд отклонений предпочтений избирателей для различных интервалов времени их расчета: (*а*) 1 день; (*б*) 10 дней

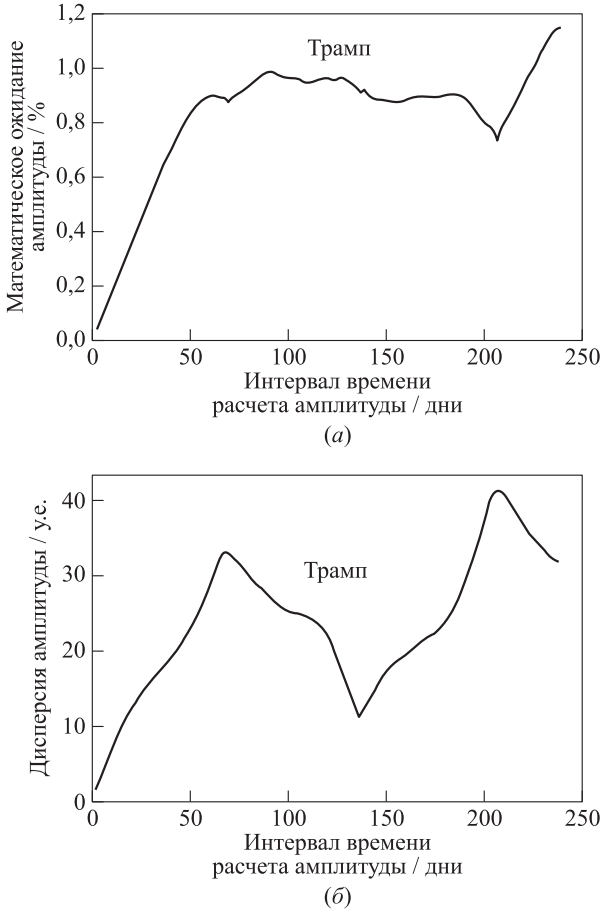


Рис. 2 Зависимости математического ожидания (*а*) и дисперсии (*б*) амплитуд отклонений предпочтений избирателей

2. Полученные для каждого из расчетных интервалов времени значения амплитуд сортируем в порядке возрастания и для каждого из интервалов строим гистограммы плотности распределения амплитуд (рис. 1).
3. По полученным гистограммам для каждого из интервалов времени расчета амплитуд вычисляем моменты распределений (среднее значение — математическое ожидание, дисперсия), проводим построение их зависимостей от величины интервала времени расчета амплитуды.

Визуальный анализ данных (см. рис. 1) показывает, что для небольших интервалов времени расчета амплитуд колебаний предпочтений избирателей (один день) гистограммы имеют большой центральный пик вблизи нулевого значения (вероятность около 0,60), а амплитуды с большими значениями имеют маленькую вероятность. При увеличении интервала времени определения амплитуды колебаний предпочтений избирателей центральный пик уменьшается, ширина распределения увеличивается, появляются осцилляции.

На рис. 2 приведены графики зависимостей математического ожидания и дисперсии амплитуд отклонений предпочтений избирателей от величины интервала времени, для которого они были рассчитаны по наблюдаемым данным (см. рис. 1). Наблюдаемые процессы носят сложный характер.

4 Модель стохастической динамики формирования состояний социальных систем с учетом процессов самоорганизации и наличия памяти

4.1 Вывод основного уравнения модели

Все множество возможных величин амплитуд отклонений предпочтений избирателей для любой величины интервала времени t обозначим X. Считаем, что интервал времени t состоит из малых частей τ. Тогда любое значение интервала времени t представим как $t_h = h\tau$, где h — номер шага ($h = 0, 1, 2, \ldots, N$). Величину амплитуды для выбранного интервала времени t обозначим x_h ($x_h \in \in$ X). Анализ наблюдаемых величин амплитуд (см. рис. 1) показывает, что x_h могут иметь положительные и отрицательные значения. Предположим, что значение амплитуды x_h при изменении дискретного времени h на единицу может увеличиваться на некоторую малую величину ε или уменьшаться на величину ξ. Найдем вероятность $\mathsf{P}(x, h)$ того, что величина амплитуды отклонений предпочтений избирателей для некоторого интервала дискретного времени h окажется равна x. Пусть $\mathsf{P}(x-\varepsilon, h-1)$ — вероятность того, что для некоторого $(h-1)$ амплитуда имела величину $(x-\varepsilon)$; $\mathsf{P}(x+\xi, h-1)$ — вероятность того, что для некоторого $(h-1)$ амплитуда имеет величину $(x+\xi)$; $\mathsf{P}(x, h-1)$ — вероятность того, что для некоторого $(h-1)$ амплитуда имеет величину x. Вероятность $\mathsf{P}(x, h)$ того, что величина амплитуды отклонения предпочтений избирателей для интервала дискретного времени h окажется равной x (рис. 3), можно определить по формуле:

$$\mathsf{P}(x, h) = \\ = \mathsf{P}(x-\varepsilon, h-1) + \mathsf{P}(x+\xi, h-1) - \mathsf{P}(x, h-1).$$

Человеческий фактор, внося неопределенность воздействием на процессы и создавая стохастичность, открывает возможности для самоорганизации и определяет наличие памяти о предыдущих

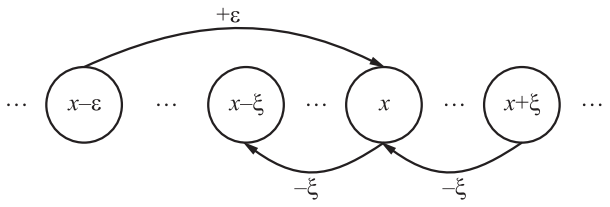


Рис. 3 Схема переходов между величинами амплитуды при изменении h на 1

действиях. Для учета памяти определим вероятности $\mathsf{P}(x-\varepsilon,h)$, $\mathsf{P}(x+\xi,h)$ и $\mathsf{P}(x,h)$ через состояния на предыдущем, $(h-1)$-м, шаге. Схемы соответствующих переходов можно изобразить аналогично схеме на рис. 3. Получаем для вероятности перехода:

$$\begin{aligned}&\mathsf{P}(x,h+2)=\\&=\{\mathsf{P}(x-2\varepsilon,h)+\mathsf{P}(x-\varepsilon+\xi,h)-\mathsf{P}(x-\varepsilon,h)\}+\\&+\{\mathsf{P}(x+\xi-\varepsilon,h)+\mathsf{P}(x+\xi,h)-\mathsf{P}(x+\xi,h)\}-\\&\quad-\mathsf{P}(x-\varepsilon,h)-\mathsf{P}(x+\xi,h-1)+\mathsf{P}(x,h)\,.\end{aligned}$$

Далее, учитывая, что $t=h\tau$, перейдем от h к t, затем разложим в ряд Тейлора:

$$\frac{d\mathsf{P}(x,t)}{dt}=a\frac{d^2\mathsf{P}(x,t)}{dx^2}-b\frac{d\mathsf{P}(x,t)}{dx}-c\frac{d^2\mathsf{P}(x,t)}{dt^2}\,,\quad(1)$$

где

$$a=\frac{\varepsilon^2-\varepsilon\xi+\xi^2}{\tau}\,;\;\; b=\frac{\varepsilon-\xi}{\tau}\,;\;\; c=\tau\,.$$

Член уравнения вида $d\mathsf{P}(x,t)/dx$ описывает упорядоченный переход либо в состояние, когда оно увеличивается ($\varepsilon>\xi$), либо когда оно уменьшается ($\varepsilon<\xi$); член уравнения вида $d^2\mathsf{P}(x,t)/dx^2$ описывает случайное изменение состояния (неопределенность изменения). Член уравнения вида $d\mathsf{P}(x,t)/dt$ определим как скорость общего изменения состояния системы с течением времени; член уравнения вида $d^2\mathsf{P}(x,t)/dt^2$ описывает процесс, при котором состояния сами становятся источниками возникновения других состояний (*самоорганизация* и ускорение как упорядоченных ($d\mathsf{P}(x,t)/dx$), так и случайных ($d^2\mathsf{P}(x,t)/dx^2$) переходов).

4.2 Формулировка и решение граничной задачи для нахождения функции распределения амплитуд отклонений предпочтений избирателей

Считая функцию $\mathsf{P}(x,t)$ непрерывной, перейдем от вероятности $\mathsf{P}(x,t)$ (1) к плотности вероятности $\rho(x,t)=d\mathsf{P}(x,t)/dx$ и сформулируем граничную задачу для нахождения зависимости плотности вероятности наблюдения различных величин амплитуд отклонений предпочтений избирателей за произвольный интервал времени t. Анализ статистических данных показывает, что вероятности наблюдения больших величин амплитуд отклонений предпочтений избирателей в течение рассматриваемых интервалов времени ничтожно малы. Можно предположить, что функция плотности вероятности быстро убывает, и задать граничные условия:

$$\rho(x,t)_{x=\infty}=0\,;\quad \rho(x,t)_{x=-\infty}=0\,.$$

Первое начальное условие зададим в виде дельта-функции, исходя из того, что для интервала времени $t=0$ возможно только значение амплитуды $x_0=0$:

$$\rho(x,t)|_{t=0}=\delta(x-0)=\begin{cases}\int\delta(x-0)\,dx=1\,, & x=0\,;\\ 0\,, & x\neq 0\,.\end{cases}$$

Второе начальное условие $(\partial\rho(x,t)/\partial t)|_{t=0}$ задает скорость изменения плотности вероятности для любого значения амплитуды. Сложение множества различных типов поведения избирателей ведет к тому, что некоторые амплитуды могут усиливаться, а некоторые ослабевать. В конечном итоге это приводит к периодичности для некоторых значений амплитуд, т. е. возникновению волн. Так как $\Delta t\to\tau$ (по условиям протекания процессов), то второе начальное условие можно записать в виде:

$$\begin{aligned}&\left.\frac{\partial\rho(x,t)}{\partial t}\right|_{t=0}=\\&=\lim_{\Delta t\to\tau}\left.\frac{\rho(x+\Delta x,t+\Delta t)-\rho(x,t)}{\Delta t}\right|_{t=0}=\\&=\frac{\rho(x+\Delta x,0+\tau)-\rho(x,0)}{\tau}=\frac{1}{\tau}\,\psi(x)\delta(x-y)\,,\end{aligned}$$

где $\psi(x)$ — некоторая периодическая функция. Из решения граничной задачи для уравнения (1) получаем для функции плотности вероятности амплитуд отклонений предпочтений избирателей следующую зависимость:

$$\begin{aligned}\rho(x,t)=&\frac{\tau U(t-k)e^{(\varepsilon-\xi)x/(2(\varepsilon^2-\varepsilon\xi+\xi^2))}e^{-t/(2\tau)}}{2\sqrt{\varepsilon^2-\varepsilon\xi+\xi^2}}\times\\&\times\left\{\frac{1}{\tau}\left\{\frac{1}{2}+\psi(x)\right\}\sum_{n=0}^{\infty}\frac{\left\{\omega^2\left\{t^2-k^2\right\}\right\}^n}{4^n(n!)^2}+\right.\\&\left.+\frac{t}{t^2-k^2}\sum_{n=0}^{\infty}\frac{2n\left\{\omega^2\left\{t^2-k^2\right\}\right\}^n}{4^n(n!)^2}\right\}\,,\quad(2)\end{aligned}$$

где

$$\omega = \sqrt{\frac{\varepsilon\xi}{4\tau^2(\varepsilon^2 - \varepsilon\xi + \xi^2)}};$$

$$k = \frac{|x|\tau}{\sqrt{\varepsilon^2 - \varepsilon\xi + \xi^2}};$$

$\psi(x)$ — периодические функции:

$$\psi(x) = \begin{cases} \cos\left\{2\pi\dfrac{x}{\lambda}\right\}; \\ \sin\left\{2\pi\dfrac{x}{\lambda}\right\}; \end{cases}$$

$U(t-k)$ — функция Хэвисайда:

$$U(t-k) = \begin{cases} 0, & \text{если } t < k; \\ \dfrac{1}{2}, & \text{если } t = k; \\ 1, & \text{если } t > k. \end{cases}$$

Величина λ в периодической функции $\psi(x)$ имеет смысл длины волны для процесса колебаний амплитуды. При отсутствии осцилляций $\psi(x) = 0$. Величина $k = |x|\tau/\sqrt{\varepsilon^2 - \varepsilon\xi + \xi^2}$ имеет смысл времени запаздывания распространения волнового процесса на время, равное k ($k = |x|/V_0$, где $V_0 = \sqrt{\varepsilon^2 - \varepsilon\xi + \xi^2}/\tau = \lambda/\tau$ — скорость распространения волнового процесса, $\lambda = \sqrt{\varepsilon^2 - \varepsilon\xi + \xi^2}$ — длина волны). Для функции $\rho(x,t)$ выполняется условие нормировки

$$\int\limits_{-\infty}^{+\infty} \rho(x,t)\,dx = 1.$$

4.3 Анализ модели, учитывающей влияние на амплитуду отклонений предпочтений избирателей, процессов самоорганизации и наличия памяти

На рис. 4 представлены результаты моделирования зависимости плотности вероятности амплитуды отклонения от интервала времени ее расчета для различных наборов параметров ε и ξ с использованием полученного уравнения (2). При увеличении длины волны $\lambda = \sqrt{\varepsilon^2 - \varepsilon\xi + \xi^2}$ число осцилляций на графиках плотности вероятности амплитуд отклонений предпочтений избирателей падает, а при уменьшении — растет. С увеличением интервала времени расчета амплитуд ширина и число осцилляций увеличиваются, а высота распределения уменьшается. При $\varepsilon > \xi$ происходит смещение максимума плотности вероятности вправо, а при $\varepsilon < \xi$ — влево. С ростом интервала времени расчета амплитуд высота распределения уменьшается, а ширина и число осцилляций увеличиваются. Наблюдаются асимметрия распределения относительно линии максимума. Если $\psi(x,\lambda) = 0$, то осцилляции исчезают, а остальные характеристики поведения распределения сохраняют свою тенденцию. При выборе другого набора параметров ξ, ε, τ и t поведение плотностей вероятности амплитуд остается прежним, но значения и положения максимумов на графиках изменяются. Сравнение наблюдаемых гистограмм распределений (см. рис. 1) с результатами теоретического моделирования (см. рис. 4) показывает хорошее соответствие разработанной модели наблюдаемым данным.

На рис. 5 представлены зависимости математического ожидания и дисперсии величины амплитуды в зависимости от интервала времени ее расчета,

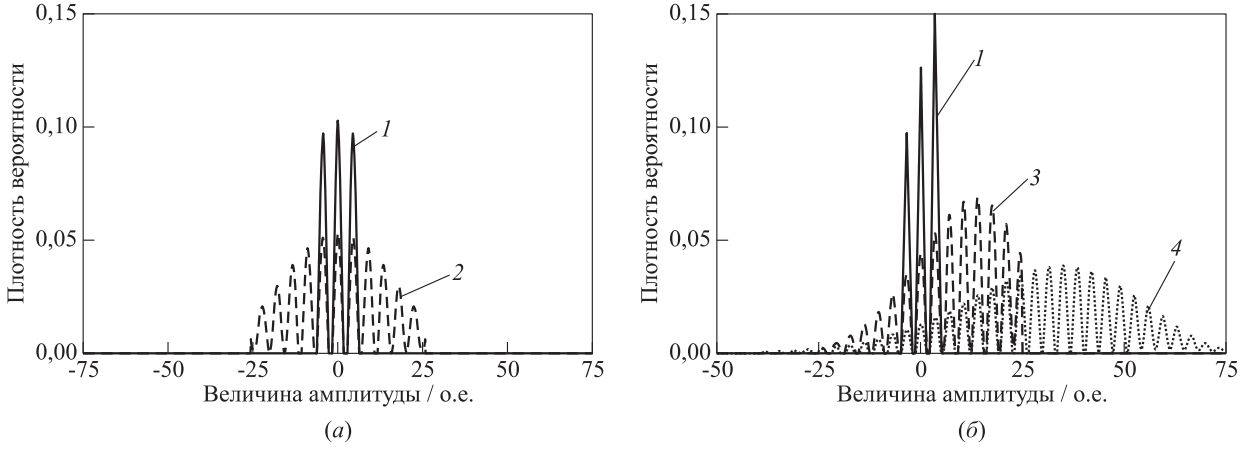


Рис. 4 Теоретические зависимости плотности вероятности величин амплитуд от времени их расчета ($\tau = 0{,}7$): (*а*) $\varepsilon = \xi = 4{,}5$; (*б*) $\varepsilon = 4{,}0$, $\xi = 2{,}5$; *1* — $t = 1$; *2* — 4; *3* — 5; *4* — $t = 15$

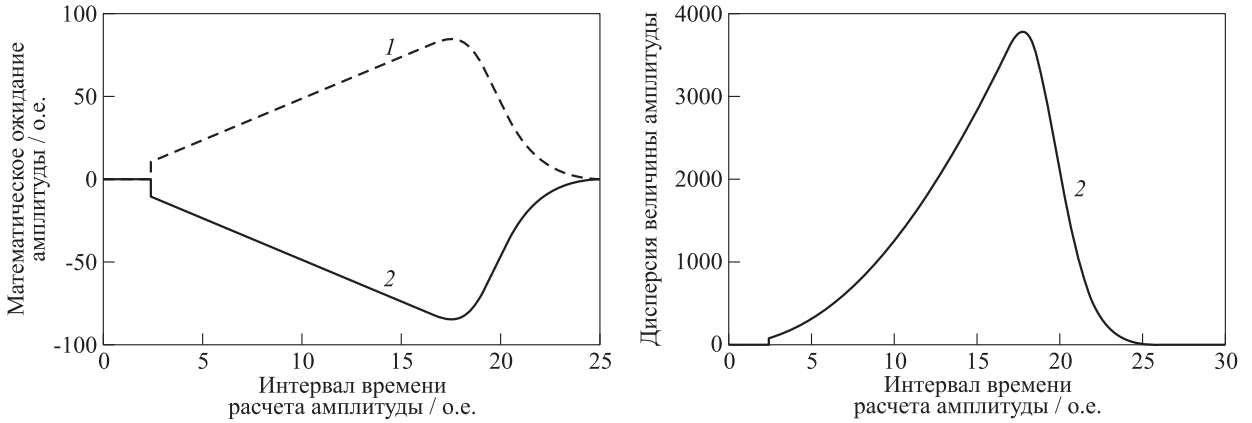


Рис. 5 Теоретические зависимости математического ожидания (*а*) и дисперсии величин амплитуд (*б*) от времени их расчета для наборов параметров ($\tau = 0{,}1$): *1* — $\xi = 0{,}2$, $\varepsilon = 0{,}7$; *2* — $\xi = 0{,}7$, $\varepsilon = 0{,}2$

полученные с использованием разработанной модели (для различных наборов величин параметров ε и ξ).

Расчеты показывают, что при $\varepsilon > \xi$ наблюдаются амплитуды роста и их значения находятся в положительной области на рис. 5 (кривая *1*). При $\varepsilon < \xi$ наблюдаются амплитуды падения, их математическое ожидание находится в отрицательной области (кривая *2*). Для наборов параметров с инверсией величин ε и ξ ($\tau = 0{,}1$): $\xi = 0{,}7$, $\varepsilon = 0{,}2$ и $\xi = 0{,}2$, $\varepsilon = 0{,}7$ — поведение дисперсии имеет одинаковый характер, так как она является квадратичной величиной и не имеет отрицательных значений.

Среднее значение амплитуды отклонения (математическое ожидание) $\mu(t)$ и дисперсия $\sigma^2(t)$ рассчитываются с использованием выражения (2) следующим образом:

$$\mu(t) = \int\limits_{-\infty}^{+\infty} x\rho(x,t)\,dx\,; \quad \sigma^2(t) = \int\limits_{-\infty}^{+\infty} x^2\rho(x,t)\,dx\,.$$

Положение максимумов и другие характеристики процессов зависят от выбора величин параметров модели: при увеличении значений параметров модели ε и ξ математическое ожидание амплитуд уменьшается, а максимум сдвигается в область малых интервалов времени. Зависимости математического ожидания и дисперсии наблюдаемых амплитуд отклонения предпочтений избирателей от величины интервала времени их расчета, полученные из социологических данных (см. рис. 2), отличаются по виду от результатов моделирования, представленных на рис. 5. Предполагая наличие нескольких процессов для выбора предпочтений избирателей по каждому из кандидатов с разными весовыми коэффициентами $\alpha_1, \alpha_2, \ldots$ и несколькими наборами параметров модели ε и ξ, получим результаты теоретического моделирования (рис. 6), хорошо соответствующие наблюдаемым данным ($\tau = 0{,}1$). Для процесса I (возрастание): $\xi_1 = 0{,}15$; $\varepsilon_1 = 0{,}45$; $\alpha_1 = 0{,}65$; для процесса II (возрастание): $\xi_2 = 0{,}20$; $\varepsilon_2 = 0{,}75$; $\alpha_2 = 0{,}15$; для процесса III (возрастание): $\xi_3 = 0{,}25$; $\varepsilon_3 = 0{,}95$; $\alpha_3 = 0{,}10$; для процесса IV (возрастание): $\xi_4 = 0{,}30$; $\varepsilon_4 = 1{,}70$; $\alpha_4 = 0{,}10$. Весовые коэффициенты каждого из процессов могут быть различными и меняться от 0 до 1. Наличие нескольких процессов для динамики предпочтений одного и того же кандидата может быть обусловлено различными группами избирателей с различным типом поведения при выборе предпочтений, а величина весовых коэффициентов может зависеть от соотношения численности каждой из групп.

Сравнение наблюдаемых данных (см. рис. 2) и результатов теоретического моделирования (см. рис. 6) показывает (с учетом приближенности моделирования), что подбором величин параметров можно получить неплохое совпадение с наблюдаемыми данными. Полученные результаты позволяют сделать общий вывод о том, что созданная модель в целом хорошо описывает наблюдаемую динамику электоральных процессов и ее можно использовать для прогнозирования.

Общий алгоритм прогнозирования:

(1) на основе наблюдаемых за какой-то промежуток времени данных (например, первая половина избирательной кампании) строим гистограммы, описывающие зависимость амплитуд отклонений предпочтений избирателей от интервала времени их расчета. Далее находим зависимости математического ожидания и дисперсии;

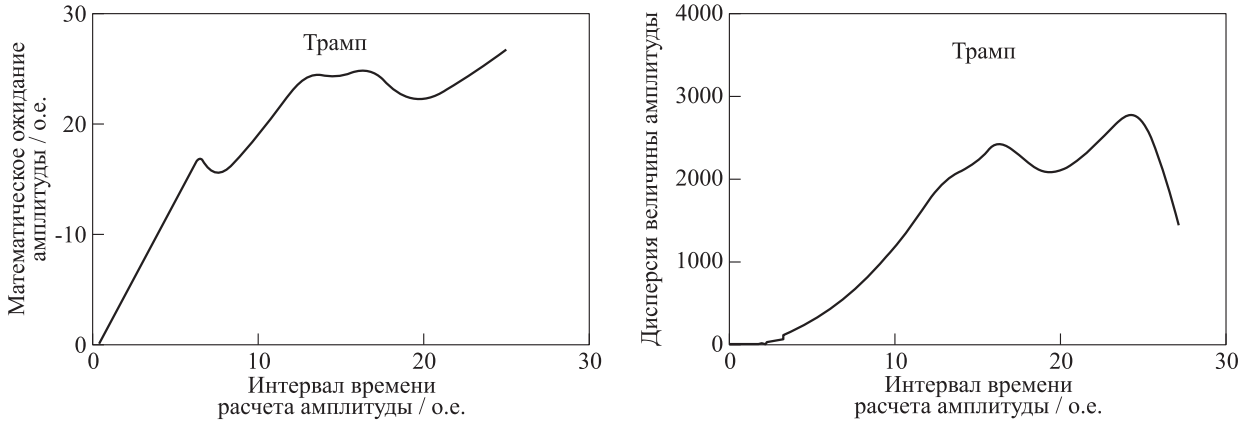


Рис. 6 Теоретическое моделирование зависимостей математического ожидания и дисперсии амплитуд отклонений предпочтений избирателей

(2) на основе уравнения (2) и наблюдаемых характеристик процессов (см. рис. 2) определяем величины параметров ξ и ε. Используем полученные результаты для расчета величины предпочтений избирателей на момент окончания избирательной кампании. Если полученный результат не приводит к победе, то необходимо повлиять на процесс и изменить его параметры. Например, за счет средств массовой информации изменить величину параметра ε в нужную сторону (увеличить свою или уменьшить у соперника).

5 Заключение и выводы

1. Проведен анализ динамики электоральной кампании по выборам президента США в 2016 г. На основании полученных данных построены гистограммы зависимостей амплитуд отклонений предпочтений избирателей от величины интервала времени их определения. На основе обработанных данных наблюдений создана модель стохастической динамики изменения предпочтения избирателей, учитывающая процессы самоорганизации, наличие памяти и хорошо описывающая основные характеристики исследуемых процессов (появление осцилляций, изменение высоты и ширины распределения при изменении интервала времени расчета амплитуд и т. д.).
2. При создании модели стохастической динамики предпочтений избирателей рассмотрены схемы вероятностей переходов между ее возможными состояниями. Выведено нелинейное дифференциальное уравнение второго порядка, сформулирована и решена граничная задача для определения функции плотности вероятности амплитуды отклонений предпочтений избирателей от величины интервала времени ее определения. Дифференциальное уравнение содержит член, отвечающий за возможность самоорганизации, а также учитывает наличие памяти.
3. Разработанную стохастическую модель динамики изменения величины предпочтений избирателей с учетом процессов самоорганизации, наличия памяти и осцилляций можно использовать для прогнозирования результатов электоральных кампаний и принятия решений.

Литература

1. *Easley D., Kleinberg J.* Networks, crowds, and markets: Reasoning about a highly connected world. — Cambridge: Cambridge University Press, 2010. 819 p. doi: 10.1017/CBO9780511761942.
2. *Karsai M., Iniguez G., Kaski K., Kertesz J.* Complex contagion process in spreading of onlne innovation // J. R. Soc. Interface, 2014. Vol. 11. Art. ID: 20140694. 8 p. doi: 10.1098/rsif.2014.0694.
3. *Gleeson J. P., Cahalane D. J.* Seed size strongly affects cascades on random networks // Phys. Rev. E, 2007. Vol. 75. Art. ID: 056103. 4 p. doi: 10.1103/PhysRevE.75.0561037.
4. *Barrat A., Barthelemy M., Vespignani A.* Dynamical processes on complex networks. — Cambridge: Cambridge University Press, 2012. 347 p. doi: 10.1017/CBO9780511791383.
5. *Kocsis G., Kun F.* Competition of information channels in the spreading of innovations // Phys. Rev. E, 2011. Vol. 84. Art. ID: 026111. 7 p. doi: 10.1103/PhysRevE.84.026111.
6. *Airoldi E. M., Blei D. M., Fienberg S. E., Xing E. P.* Mixed membership stochastic blockmodels // J. Mach. Learn. Res., 2008. Vol. 9. P. 1981–2014.

7. *Khvatova T., Block M., Zhukov D., Lesko S.* How to measure trust: The percolation model applied to intra-organisational knowledge sharing networks // J. Knowl. Manag., 2016. Vol. 20. Iss. 5. P. 918–935. doi: 10.1108/JKM-11-2015-0464.
8. *Khvatova T. Yu., Zaltsman A. D., Zhukov D. O.* Information processes in social networks: Percolation and stochastic dynamics // CEUR Workshop Procee., 2017. Vol. 2064. P. 277–288.
9. *Plikynas D., Raudys A., Raudys S.* Agent-based modelling of excitation propagation in social media groups // J. Experimental Theoretical Artificial Intelligence, 2015. Vol. 27. Iss. 4. P. 373–388. doi: 10.1080/0952813X.2014.954631.
10. *Wang A., Wu W., Chen J.* Social network rumors spread model based on cellular automata // 10th Conference (International) on Mobile Ad-hoc and Sensor Networks Proceedings. — Piscataway, NJ, USA: IEEE, 2014. P. 236–242. doi: 10.1109/MSN.2014.39.
11. *Андрианова Е. Г., Головин С. А., Зыков С. В., Лесько С. А., Чукалина Е. Р.* Обзор современных моделей и методов анализа временных рядов динамики процессов в социальных, экономических и социотехнических системах // Российский технологический ж., 2020. Т. 8. № 4(36). С. 7–45. doi: 10.32362/2500-316X-2020-8-4-7-45.
12. *Zhukov D. O., Lesko S. A., Khvatova T. Yu.* Percolation models of information distribution and blocking in social networks // 5th Ashridge Research Conference (International) Global Disruption and Organisational Innovation. — Berkhamsted, U.K., 2016. Art. ID: 23423.
13. *Zhukov D., Khvatova T., Zaltsman A.* Stochastic dynamics of influence expansion in social networks and managing users' transitions from one state to another // 11th European Conference on Information Systems Management Proceedings. — Reading: Academic Conferences and Publishing International Ltd., 2017. P. 322–329.
14. *Zhukov D. O., Alyoshkin A. S., Obukhova A. G.* Modelling to be based on systems of differential kinetic equations to processes group selection voters during the electoral campaign of Trump–Clinton 2015–2016 // 7th Conference (International) on Information Communication and Management Proceedings. — New York, NY, USA: ACM, 2017. P. 88–94.
15. *Sigov A. S., Zhukov D. O., Khvatova T. Yu., Andrianova E. G.* Model of forecasting of information events on the basis of the solution of a boundary value problem for systems with memory and self-organization // J. Commun. Technol. El., 2018. Vol. 18. Iss. 2. P. 106–117.
16. *Zhukov D., Khvatova T., Istratov L.* A stochastic dynamics model for shaping stock indexes using self-organization processes, memory and oscillations // European Conference on the Impact of Artificial Intelligence and Robotics Proceedings. — Oxford, U.K.: ACPIL, 2019. P. 390–401.
17. *Zhukov D., Zaltsman A., Khvatova T.* Forecasting changes in states in social networks and sentiment security using the principles of percolation theory and stochastic dynamics // IEEE Conference (International) "Quality Management, Transport and Information Security, Information Technologies." — Piscataway, NJ, USA: IEEE, 2019. P. 149–153.
18. *Smychkova A., Zhukov D.* Complex of description models for analysis and control group behavior based on stochastic cellular automata with memory and systems of differential kinetic equations // 1st Conference (International) on Control Systems, Mathematical Modelling, Automation and Energy Efficiency Proceedings. — Lipetsk: Lipetsk State Technical University, 2019. P. 218–223.
19. *Zhukov D., Khvatova T., Millar C., Zaltsman A.* Modeling the stochastic dynamics of influence expansion and managing transitions between states in social networks // Technol. Forecast. Soc., 2020. Vol. 158. P. 1–15.
20. *Жуков Д. О., Хватова Т. Ю., Зальцман А. Д.* Моделирование стохастической динамики изменения состояний узлов и перколяционных переходов в социальных сетях с учетом самоорганизации и наличия памяти // Информатика и её применения, 2021. Т. 15. Вып. 1. С. 102–110.
21. *Zhukov D., Andrianova E., Trifonova O.* Stochastic diffusion model for analysis of dynamics and forecasting events in news feeds // Symmetry, 2021. Vol. 13. Iss. 2. Art. No. 257. 21 p. doi: 10.3390/sym13020257.
22. *Zhukov D., Andrianova E., Novikova O.* Diffusion model for forecasting events in news feeds // J. Phys. Conf. Ser., 2021. Vol. 1727. Iss. 1. P. 21–32.

Поступила в редакцию 15.09.2019

STOCHASTIC DYNAMICS OF SELF-ORGANIZING SOCIAL SYSTEMS WITH MEMORY (ELECTORAL PROCESSES)

A. S. Sigov, E. G. Andrianova, and L. A. Istratov

Russian Technological University (MIREA), 78 Vernadskogo Ave., Moscow 119454, Russian Federation

Abstract: The paper discusses the use of the methods and approaches which are common for theoretical computer science as well as the use of its applications for analysis and modeling of social group processes. Based on the developed model for describing stochastic processes, taking into account self-organization and the presence of memory, an analysis of the voter preference dynamics during the 2016 U.S. presidential campaign was conducted.

The sociological data processing allowed plotting the probability density histograms for the amplitudes of voter preference deviation, depending on their determination interval, and developing a model that well describes the main characteristics of the observed processes (appearance of oscillations, changes in the height and width of the distribution depending on the changes in the amplitude calculation interval, etc.). In the course of building the model, the probability schemes of transitions between the possible states of the social system (voter preferences) were considered and a second-order nonlinear differential equation was derived. In addition, a boundary problem to determine the probability density function of the amplitude of voter preference deviation depending on its determination interval was formulated and solved. The model differential equation has a term responsible for the self-organization possibility and takes into account the presence of memory. The oscillation possibility depends on the initial conditions. The developed model can be used for analyzing election campaigns and making relevant decisions.

Keywords: oscillation amplitude distribution function; stochastic dynamics; self-organization; presence of memory; probability density oscillations; electoral processes

DOI: 10.14357/19922264210216

References

1. Easley, D., and J. Kleinberg. 2010. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge: Cambridge University Press. 819 p. doi: 10.1017/CBO9780511761942.
2. Karsai, M., G. Iniguez, K. Kaski, and J. Kertesz. 2014. Complex contagion process in spreading of online innovation. *J. R. Soc. Interface* 11:20140694. 8 p. doi: 10.1098/rsif.2014.0694.
3. Gleeson, J. P., and D. J. Cahalane. 2007. Seed size strongly affects cascades on random networks. *Phys. Rev. E* 75:056103. 4 p. doi: 10.1103/PhysRevE.75.0561037.
4. Barrat, A., M. Barthelemy, and A. Vespignani. 2012. *Dynamical processes on complex networks*. Cambridge: Cambridge University Press. 347 p. doi: 10.1017/CBO9780511791383.
5. Kocsis, G., and F. Kun. 2011. Competition of information channels in the spreading of innovations. *Phys. Rev. E* 84:026111. 7 p. doi: 10.1103/PhysRevE.84.026111.
6. Airoldi, E. M., D. M. Blei, S. E. Fienberg, and E. P. Xing. 2008. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* 9:1981–2014.
7. Khvatova, T., M. Block, D. Zhukov, and S. Lesko. 2016. How to measure trust: The percolation model applied to intraorganisational knowledge sharing networks. *J. Knowl. Manag.* 20(5):918–935. doi: 10.1108/JKM-11-2015-0464.
8. Khvatova, T. Yu., A. D. Zaltcman, and D. O. Zhukov. 2017. Information processes in social networks: Percolation and stochastic dynamics. *CEUR Workshop Procee*. 2064:277–288.
9. Plikynas, D., A. Raudys, and S. Raudys. 2015. Agent-based modelling of excitation propagation in social media groups. *J. Exp. Theor. Artif. In.* 27(4):373–388. doi: 10.1080/0952813X.2014.954631.
10. Wang, A., W. Wu, and J. Chen. 2014. Social network rumors spread model based on cellular automata. *10th Conference (International) on Mobile Ad-Hoc and Sensor Networks Proceedings*. Piscatway, NJ: IEEE. 236–242. doi: 10.1109/MSN.2014.39.
11. Andrianova, E. G., S. A. Golovin, S. V. Zykov, S. A. Lesko, and E. R. Chukalina. 2020. Obzor sovremennykh modeley i metodov analiza vremennykh ryadov dinamiki protsessov v sotsial'nykh, ekonomicheskikh i sotsiotekhnicheskikh sistemakh [Review of modern models and methods of analysis of time series of dynamics of processes in social, economic and socio-technical systems]. *Russ. Technological J.* 8(4):7–45. doi: 10.32362/2500-316X-2020-8-4-7-45.
12. Zhukov, D. O., S. A. Lesko, and T. Yu. Khvatova. 2016. Percolation models of information distribution and blocking in social networks. *5th Ashridge Research Conference (International) Global Disruption and Organisational Innovation Proceedings*. Berkhamsted, U.K. 23423.
13. Zhukov, D., T. Khvatova, and A. Zaltsman. 2017. Stochastic dynamics of influence expansion in social networks and managing users' transitions from one state to another. *11th European Conference on Information Systems Management Proceedings*. Reading: Academic Publishing International Ltd. 322–329.
14. Zhukov, D. O., A. S. Alyoshkin, and A. G. Obukhova. 2017. Modelling to be based on systems of differential kinetic equations to processes group selection voters during the electoral campaign of Trump–Clinton 2015–2016. *7th Conference (International) on Information Communication and Management Proceedings*. New York, NY: ACM. 88–94.
15. Sigov, A. S., D. O. Zhukov, T. Yu. Khvatova, and E. G. Andrianova. 2018. Model of forecasting of information events on the basis of the solution of a boundary value problem for systems with memory and self-organization. *J. Commun. Technol. El.* 18(2):106–117.
16. Zhukov, D., T. Khvatova, and L. Istratov. 2019. A stochastic dynamics model for shaping stock indexes using self-organization processes, memory and oscillations. *European Conference on the Impact of Artificial Intelligence and Robotics Proceedings*. Oxford, U.K.: ACPIL. 390–401.
17. Zhukov, D., A. Zaltsman, and T. Khvatova. 2019. Forecasting changes in states in social networks and sentiment security using the principles of percolation theory and stochastic dynamics. *Conference (International) "Quality Management, Transport and Information Security, Infor-*

mation Technologies" Proceedings. Piscataway, NJ: IEEE. 149–153.

18. Smychkova, A., and D. Zhukov. 2019. Complex of description models for analysis and control group behavior based on stochastic cellular automata with memory and systems of differential kinetic equations. *Ist Conference (International) on Control Systems, Mathematical Modelling, Automation and Energy Efficiency Proceedings*. Lipetsk: Lipetsk State Technical University. 218–223.
19. Zhukov, D., T. Khvatova, C. Millar, and A. Zaltsman. 2020. Modeling the stochastic dynamics of influence expansion and managing transitions between states in social networks. *Technol. Forecast. Soc*. 158:1–15.
20. Zhukov, D. O., T. Yu. Khvatova, and A. D. Zaltcman. 2021. Modelirovanie stokhasticheskoy dinamiki izmereniya sostoyaniy uzlov i perkolyatsionnykh perekhodov v sotsial'nykh setyakh s uchetom samoorganizatsii i nalichiya pamyati [Modeling of the stochastic dynamics of changes in node states and percolation transitions in social networks with self-organization and memory]. *Informatika i ee Primeneniya — Inform. Appl*. 15(1):102–110.
21. Zhukov, D., E. Andrianova, and O. Trifonova. 2021. Stochastic diffusion model for analysis of dynamics and forecasting events in news feeds. *Symmetry* 13(2):257. 21 p. doi: 10.3390/sym13020257.
22. Zhukov, D., E. Andrianova, and O. Novikova. 2021. Diffusion model for forecasting events in news feeds. *J. Phys. Conf. Ser.* 1727(1):21–32.

Received September 15, 2019

Contributors

Sigov Alexander S. (b. 1945) — Doctor of Science in physics and mathematics, professor, Academician of RAS, President of the Russian Technological University (MIREA), 78 Vernadskogo Ave., Moscow 119454, Russian Federation; sigov@mirea.ru

Andrianova Elena G. (b. 1963) — Candidate of Science (PhD) in technology, associate professor, Russian Technological University (MIREA), 78 Vernadskogo Ave., Moscow 119454, Russian Federation; andrianova@mirea.ru

Istratov Leonid A. (b. 1991) — student, Russian Technological University (MIREA), 78 Vernadskogo Ave., Moscow 119454, Russian Federation; istratov@mirea.ru

ОБ АВТОРАХ

Андрианова Елена Гельевна (р. 1963) — кандидат технических наук, доцент МИРЭА — Российского технологического университета

Базилевский Михаил Павлович (р. 1987) — кандидат технических наук, доцент Иркутского государственного университета путей сообщения

Борисов Андрей Владимирович (р. 1965) — доктор физико-математических наук, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; профессор Московского авиационного института; профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; старший научный сотрудник Московского центра фундаментальной и прикладной математики Московского государственного университета имени М. В. Ломоносова

Босов Алексей Вячеславович (р. 1969) — доктор технических наук, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Гончаренко Мирослав Богданович (р. 1991) — старший специалист по разработке программного обеспечения в области компьютерной графики АО «Интел А/О»

Гончаров Александр Анатольевич (р. 1994) — младший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Грушо Александр Александрович (р. 1946) — доктор физико-математических наук, профессор, главный научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Егорова Анна Юрьевна (р. 1991) — младший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Забежайло Михаил Иванович (р. 1956) — доктор физико-математических наук, доцент, главный научный сотрудник Вычислительного центра им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

Захарова Татьяна Валерьевна (р. 1962) — кандидат физико-математических наук, доцент кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Зацман Игорь Моисеевич (р. 1952) — доктор технических наук, заведующий отделом Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Зейфман Александр Израилевич (р. 1954) — доктор физико-математических наук, профессор, заведующий кафедрой Вологодского государственного университета; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; главный научный сотрудник Вологодского научного центра Российской академии наук

Инькова Ольга Юрьевна (р. 1965) — доктор филологических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Истратов Леонид Анатольевич (р. 1991) — студент МИРЭА — Российского технологического университета

Казанчян Драстамат Хачатурович (р. 1994) — аспирант кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова

Кириков Игорь Александрович (р. 1955) — кандидат технических наук, директор Калининградского филиала Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ковалёв Иван Александрович (р. 1961) — аспирант кафедры прикладной математики Вологодского государственного университета

Кривенко Михаил Петрович (р. 1946) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Листопад Сергей Викторович (р. 1984) — кандидат технических наук, старший научный сотрудник Калининградского филиала Федерального исследовательского центра «Информатика и управление» Российской академии наук

Монахов Михаил Михайлович (р. 1993) — лаборант Московского центра фундаментальной и прикладной математики Московского государственного университета имени М. В. Ломоносова

Нуриев Виталий Александрович (р. 1980) — кандидат филологических наук, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Сатин Яков Александрович (р. 1978) — кандидат физико-математических наук, доцент кафедры математики и информатики Вологодского государственного университета

Сигов Александр Сергеевич (р. 1945) — доктор физико-математических наук, профессор, академик РАН, президент МИРЭА — Российского технологического университета

Смирнов Дмитрий Владимирович (р. 1984) — бизнес-партнер по ИТ департамента безопасности ПАО «Сбербанк России»

Сушко Дмитрий Викторович (р. 1962) — кандидат физико-математических наук, старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Тимонина Елена Евгеньевна (р. 1952) — доктор технических наук, профессор, ведущий научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ушаков Владимир Георгиевич (р. 1952) — доктор физико-математических наук, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Ушаков Николай Георгиевич (р. 1954) — доктор физико-математических наук, ведущий научный сотрудник Института проблем технологии микроэлектроники и особо чистых материалов Российской академии наук; профессор Норвежского научно-технологического университета

Шестаков Олег Владимирович (р. 1976) — доктор физико-математических, профессор кафедры математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова; старший научный сотрудник Института проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук

Правила подготовки рукописей для публикации в журнале «Информатика и её применения»

Журнал «Информатика и её применения» публикует теоретические, обзорные и дискуссионные статьи, посвященные научным исследованиям и разработкам в области информатики и ее приложений.

Журнал издается на русском языке. По специальному решению редколлегии отдельные статьи могут печататься на английском языке.

Тематика журнала охватывает следующие направления:

- теоретические основы информатики;
- математические методы исследования сложных систем и процессов;
- информационные системы и сети;
- информационные технологии;
- архитектура и программное обеспечение вычислительных комплексов и сетей.

1. В журнале печатаются статьи, содержащие результаты, ранее не опубликованные и не предназначенные к одновременной публикации в других изданиях.

   Публикация предоставленной автором(ами) рукописи не должна нарушать положений глав 69, 70 раздела VII части IV Гражданского кодекса, которые определяют права на результаты интеллектуальной деятельности и средства индивидуализации, в том числе авторские права, в РФ.

   Ответственность за нарушение авторских прав, в случае предъявления претензий к редакции журнала, несут авторы статей.

   Направляя рукопись в редакцию, авторы сохраняют свои права на данную рукопись и при этом передают учредителям и редколлегии журнала неисключительные права на издание статьи на русском языке (или на языке статьи, если он отличен от русского) и на перевод ее на английский язык, а также на ее распространение в России и за рубежом. Каждый автор должен представить в редакцию подписанный с его стороны «Лицензионный договор о передаче неисключительных прав на использование произведения», текст которого размещен по адресу http://www.ipiran.ru/publications/licence.doc. Этот договор может быть представлен в бумажном (в 2-х экз.) или в электронном виде (отсканированная копия заполненного и подписанного документа).

   Редколлегия вправе запросить у авторов экспертное заключение о возможности публикации представленной статьи в открытой печати.

2. К статье прилагаются данные автора (авторов) (см. п. 8). При наличии нескольких авторов указывается фамилия автора, ответственного за переписку с редакцией.

3. Редакция журнала осуществляет экспертизу присланных статей в соответствии с принятой в журнале процедурой рецензирования.

   Возвращение рукописи на доработку не означает ее принятия к печати.

   Доработанный вариант с ответом на замечания рецензента необходимо прислать в редакцию.

4. Решение редколлегии о публикации статьи или ее отклонении сообщается авторам.

   Редколлегия может также направить авторам текст рецензии на их статью. Дискуссия по поводу отклоненных статей не ведется.

5. Редактура статей высылается авторам для просмотра. Замечания к редактуре должны быть присланы авторами в кратчайшие сроки.

6. Рукопись предоставляется в электронном виде в форматах MS WORD (.doc или .docx) или LaTeX (.tex), дополнительно — в формате .pdf, на дискете, лазерном диске или электронной почтой. Предоставление бумажной рукописи необязательно.

7. При подготовке рукописи в MS Word рекомендуется использовать следующие настройки.

   Параметры страницы: формат — A4; ориентация — книжная; поля (см): внутри — 2,5, снаружи — 1,5, сверху — 2, снизу — 2, от края до нижнего колонтитула — 1,3.

   Основной текст: стиль — «Обычный», шрифт — Times New Roman, размер — 14 пунктов, абзацный отступ — 0,5 см, 1,5 интервала, выравнивание — по ширине.

Рекомендуемый объем рукописи — не свыше 10 страниц указанного формата. При превышении указанного объема редколлегия вправе потребовать от автора сокращения объема рукописи.

Сокращения слов, помимо стандартных, не допускаются. Допускается минимальное количество аббревиатур.

Все страницы рукописи нумеруются.

Шаблоны примеров оформления представлены в Интернете: http://www.ipiran.ru/journal/template.doc

8. Статья должна содержать следующую информацию на ***русском и английском языках***:
   - название статьи;
   - Ф.И.О. авторов, на английском можно только имя и фамилию;
   - место работы, с указанием почтового адреса организации и электронного адреса каждого автора;
   - сведения об авторах, в соответствии с форматом, образцы которого представлены на страницах: http://www.ipiran.ru/journal/issues/2013_07_01/authors.asp и http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;
   - аннотация (не менее 100 слов на каждом из языков). Аннотация — это краткое резюме работы, которое может публиковаться отдельно. Она является основным источником информации в информационных системах и базах данных. Английская аннотация должна быть оригинальной, может не быть дословным переводом русского текста и должна быть написана хорошим английским языком. В аннотации не должно быть ссылок на литературу и, по возможности, формул;
   - ключевые слова — желательно из принятых в мировой научно-технической литературе тематических тезаурусов. Предложения не могут быть ключевыми словами;
   - источники финансирования работы (ссылки на гранты, проекты, поддерживающие организации и т. п.).

9. Требования к спискам литературы.

   Ссылки на литературу в тексте статьи нумеруются (в квадратных скобках) и располагаются в каждом из списков литературы в порядке первых упоминаний.

   Списки литературы представляются в двух вариантах:

   (1) **Список литературы к русскоязычной части**. Русские и английские работы — на языке и в алфавите оригинала;

   (2) **References**. Русские работы и работы на других языках — в латинской транслитерации с переводом на английский язык; английские работы и работы на других языках — на языке оригинала.

   Необходимо для составления списка “References” пользоваться размещенной на сайте http://www.translit.net/ru/bgn/ бесплатной программой транслитерации русского текста в латиницу.

   Список литературы “References” приводится полностью отдельным блоком, повторяя все позиции из списка литературы к русскоязычной части, независимо от того, имеются или нет в нем иностранные источники. Если в списке литературы к русскоязычной части есть ссылки на иностранные публикации, набранные латиницей, они полностью повторяются в списке “References”.

   Ниже приведены примеры ссылок на различные виды публикаций в списке “References”.

   Описание статьи из журнала:

   Zagurenko, A. G., V. A. Korotovskikh, A. A. Kolesnikov, A. V. Timonov, and D. V. Kardymon. 2008. Tekhniko-ekonomicheskaya optimizatsiya dizayna gidrorazryva plasta [Technical and economic optimization of the design of hydraulic fracturing]. *Neftyanoe hozyaystvo* [*Oil Industry*] 11:54–57.

   Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Russ. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

   Описание статьи из электронного журнала:

   Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

   Описание статьи из продолжающегося издания (сборника трудов):

   Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental’noe issledovanie prochnosti soedineniy “stal’–kompozit” [Experimental study of the strength of joints “steel–composite”]. *Trudy MGTU “Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem”* [*Bauman MSTU “Mathematical Modeling of Complex Technical Systems” Proceedings*]. 593:125–130.

Описание материалов конференций:

Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

Описание книги (монографии, сборники):

Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

Latyshev, V. N. 2009. *Tribologiya rezaniya. Kn. 1: Friktsionnye protsessy pri rezanii metallov* [*Tribology of cutting. Vol. 1: Frictional processes in metal cutting*]. Ivanovo: Ivanovskii State Univ. 108 p.

Описание переводной книги (в списке литературы к русскоязычной части необходимо указать: / Пер. с англ. — после названия книги, а в конце ссылки указать оригинал книги в круглых скобках):

1. В русскоязычной части:
   Тимошенко С. П., Янг Д. Х., Уивер У. Колебания в инженерном деле / Пер. с англ. — М.: Машиностроение, 1985. 472 с. (*Timoshenko S. P., Young D. H., Weaver W.* Vibration problems in engineering. — 4th ed. — New York, NY, USA: Wiley, 1974. 521 p.)
2. В англоязычной части:
   Timoshenko, S. P., D. H. Young, and W. Weaver. 1974. *Vibration problems in engineering*. 4th ed. New York: Wiley. 521 p.

Описание неопубликованного документа:

Latypov, A. R., M. M. Khasanov, and V. A. Baikov. 2004 (unpubl.). Geologiya i dobycha (NGT GiD) [Geology and production (NGT GiD)]. Certificate on official registration of the computer program No. 2004611198.

Описание интернет-ресурса:

Pravila tsitirovaniya istochnikov [Rules for the citing of sources]. Available at: http://www.scribd.com/doc/1034528/ (accessed February 7, 2011).

Описание диссертации или автореферата диссертации:

Semenov, V. I. 2003. Matematicheskoe modelirovanie plazmy v sisteme kompaktnyy tor [Mathematical modeling of the plasma in the compact torus]. Moscow. D.Sc. Diss. 272 p.

Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. Moscow: IPI RAN. PhD Thesis. 23 p.

Описание ГОСТа:

GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. Moscow: Standardinform Publs. 10 p.

Описание патента:

Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

10. Присланные в редакцию материалы авторам не возвращаются.
11. При отправке файлов по электронной почте просим придерживаться следующих правил:
    - указывать в поле subject (тема) название журнала и фамилию автора;
    - использовать attach (присоединение);
    - в состав электронной версии статьи должны входить: файл, содержащий текст статьи, и файл(ы), содержащий(е) иллюстрации.
12. Журнал «Информатика и её применения» является некоммерческим изданием. Плата за публикацию не взимается, гонорар авторам не выплачивается.

Адрес редакции журнала «Информатика и её применения»:
Москва 119333, ул. Вавилова, д. 44, корп. 2, ФИЦ ИУ РАН
Тел.: +7 (499) 135-86-92 Факс: +7 (495) 930-45-05
e-mail: iiep@frccsc.ru (Стригина Светлана Николаевна)
http://www.ipiran.ru/journal/issues/

Requirements for manuscripts submitted to Journal "Informatics and Applications"

Journal "Informatics and Applications" (Inform. Appl.) publishes theoretical, review, and discussion articles on the research and development in the field of informatics and its applications.

The journal is published in Russian. By a special decision of the editorial board, some articles can be published in English.

The topics covered include the following areas:

- theoretical fundamentals of informatics;
- mathematical methods for studying complex systems and processes;
- information systems and networks;
- information technologies; and
- architecture and software of computational complexes and networks.

1. The Journal publishes original articles which have not been published before and are not intended for simultaneous publication in other editions. An article submitted to the Journal must not violate the Copyright law. Sending the manuscript to the Editorial Board, the authors retain all rights of the owners of the manuscript and transfer the nonexclusive rights to publish the article in Russian (or the language of the article, if not Russian) and its distribution in Russia and abroad to the Founders and the Editorial Board. Authors should submit a letter to the Editorial Board in the following form:

   Agreement on the transfer of rights to publish:

   "*We, the undersigned authors of the manuscript ". . . ", pass to the Founder and the Editorial Board of the Journal "Informatics and Applications" the nonexclusive right to publish the manuscript of the article in Russian (or in English) in both print and electronic versions of the Journal. We affirm that this publication does not violate the Copyright of other persons or organizations.*

   Author(s) signature(s): (name(s), address(es), date).

   This agreement should be submitted in paper form or in the form of a scanned copy (signed by the authors).

2. A submitted article should be attached with **the data on the author(s)** (see item 8). If there are several authors, the contact person should be indicated who is responsible for correspondence with the Editorial Board and other authors about revisions and final approval of the proofs.
3. The Editorial Board of the Journal examines the article according to the established reviewing procedure. If the authors receive their article for correction after reviewing, it does not mean that the article is approved for publication. The corrected article should be sent to the Editorial Board for the subsequent review and approval.
4. The decision on the article publication or its rejection is communicated to the authors. The Editorial Board may also send the reviews on the submitted articles to the authors. Any discussion upon the rejected articles is not possible.
5. The edited articles will be sent to the authors for proofread. The comments of the authors to the edited text of the article should be sent to the Editorial Board as soon as possible.
6. The manuscript of the article should be presented electronically in the MS WORD (.doc or .docx) or LaTeX (.tex) formats, and additionally in the .pdf format. All documents may be sent by e-mail or provided on a CD or diskette. A hard copy submission is not necessary.
7. The recommended typesetting instructions for manuscript.

   Pages parameters: format A4, portrait orientation, document margins (cm): left — 2.5, right — 1.5, above — 2.0, below — 2.0, footer 1.3.

   Text: font —Times New Roman, font size — 14, paragraph indent — 0.5, line spacing — 1.5, justified alignment.

   The recommended manuscript size: not more than 10 pages of the specified format. If the specified size exceeded, the editorial board is entitled to require the author to reduce the manuscript.

   Use only standard abbreviations. Avoid abbreviations in the title and abstract. The full term for which an abbreviation stands should precede its first use in the text unless it is a standard unit of measurement.

   All pages of the manuscript should be numbered.

   The templates for the manuscript typesetting are presented on site: http://www.ipiran.ru/journal/template.doc.
8. The articles should enclose data both in **Russian and English**:

   - title;
   - author's name and surname;
   - affiliation — organization, its address with ZIP code, city, country, and official e-mail address;
   - data on authors according to the format: (see site)
     http://www.ipiran.ru/journal/issues/2013_07_01/authors.asp and
     http://www.ipiran.ru/journal/issues/2013_07_01_eng/authors.asp;

- abstract (not less than 100 words) both in Russian and in English. Abstract is a short summary of the article that can be published separately. The abstract is the main source of information on the article and it could be included in leading information systems and data bases. The abstract in English has to be an original text and should not be an exact translation of the Russian one. Good English is required. In abstracts, avoid references and formulae;
- indexing is performed on the basis of keywords. The use of keywords from the internationally accepted thematic Thesauri is recommended.

  Important! Keywords must not be sentences;
- Acknowledgments.

9. References. Russian references have to be presented both in English translation and Latin transliteration (refer http://www.translit.net/ru/bgn/).

   Please take into account the following examples of Russian references appearance:

   Article in journal:

   Zhang, Z., and D. Zhu. 2008. Experimental research on the localized electrochemical micromachining. *Russ. J. Electrochem.* 44(8):926–930. doi:10.1134/S1023193508080077.

   Journal article in electronic format:

   Swaminathan, V., E. Lepkoswka-White, and B. P. Rao. 1999. Browsers or buyers in cyberspace? An investigation of electronic factors influencing electronic exchange. *JCMC* 5(2). Available at: http://www.ascusc.org/jcmc/vol5/issue2/ (accessed April 28, 2011).

   Article from the continuing publication (collection of works, proceedings):

   Astakhov, M. V., and T. V. Tagantsev. 2006. Eksperimental'noe issledovanie prochnosti soedineniy "stal'–kompozit" [Experimental study of the strength of joints "steel–composite"]. *Trudy MGTU "Matematicheskoe modelirovanie slozhnykh tekhnicheskikh sistem"* [*Bauman MSTU "Mathematical Modeling of Complex Technical Systems" Proceedings*]. 593:125–130.

   Conference proceedings:

   Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Ju. Muhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *Trudy 6-go Mezhdunarodnogo Simpoziuma "Novye resursosberegayushchie tekhnologii nedropol'zovaniya i povysheniya neftegazootdachi"* [*6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*]. Moscow. 267–272.

   Books and other monographs:

   Lindorf, L. S., and L. G. Mamikoniants, eds. 1972. *Ekspluatatsiya turbogeneratorov s neposredstvennym okhlazhdeniem* [*Operation of turbine generators with direct cooling*]. Moscow: Energy Publs. 352 p.

   Dissertation and Thesis:

   Kozhunova, O. S. 2009. Tekhnologiya razrabotki semanticheskogo slovarya informatsionnogo monitoringa [Technology of development of semantic dictionary of information monitoring system]. Moscow: IPI RAN. PhD Thesis. 23 p.

   State standards and patents:

   GOST 8.586.5-2005. 2007. Metodika vypolneniya izmereniy. Izmerenie raskhoda i kolichestva zhidkostey i gazov s pomoshch'yu standartnykh suzhayushchikh ustroystv [Method of measurement. Measurement of flow rate and volume of liquids and gases by means of orifice devices]. M.: Standardinform Publs. 10 p.

   Bolshakov, M. V., A. V. Kulakov, A. N. Lavrenov, and M. V. Palkin. 2006. Sposob orientirovaniya po krenu letatel'nogo apparata s opticheskoy golovkoy samonavedeniya [The way to orient on the roll of aircraft with optical homing head]. Patent RF No. 2280590.

   References in Latin transcription are presented in the original language.

   References in the text are numbered according to the order of their first appearance; the number is placed in square brackets. All items from the reference list should be cited.

10. Manuscripts and additional materials are not returned to Authors by the Editorial Board.
11. Submissions of files by e-mail must include:
    - the journal title and author's name in the "Subject" field;
    - an article and additional materials have to be attached using the "attach" function;
    - an electronic version of the article should contain the file with the text and a separate file with figures.
12. "Informatics and Applications" journal is not a profit publication. There are no charges for the authors as well as there are no royalties.

Editorial Board address:
FRC CSC RAS, 44, block 2, Vavilov Str., Moscow 119333, Russia
Ph.: +7 (499) 135 86 92, Fax: +7 (495) 930 45 05
e-mail: iiep@frccsc.ru (to Svetlana Strigina)
http://www.ipiran.ru/english/journal.asp