Research on ocean data analysis model based on Canopy-FFCM clustering algorithm

Haiyang WANG[1,2], Mengxing HUANG[1,2,*], Di WU[1,3,*], Hao Wang[4], Miao ZHU[1,2], Haotian HE[1,2], Jian KANG[1,2]

[1] State Key Laboratory Marine Resource Utilization in South China Sea

[2] The College of Information and communication Engineering, Hainan University

[3] The College of Computer and Cyberspace Security, Hainan University

[4] Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology Haikou, China

* Mengxing Huang and Di Wu are the corresponding authors (E-mail:huangmx09@163.com, hainuwudi@163.com)

***Abstract*—With the further advancement of the research on marine environmental science, more and more attention has been paid to modern data monitoring technology. In order to obtain the current water quality information quickly, this paper combines the advantages of Canopy algorithm and FFCM algorithm, and proposes an improved Canopy-FFCM clustering algorithm. The algorithm firstly uses Canopy algorithm to quickly obtain the best clustering number, and then iterates through FFCM algorithm. Through the use of FCM algorithm, FFCM algorithm and improved Canopy-FFCM algorithm to conduct simulation analysis on samples for many times respectively, the experimental results show that compared with the traditional algorithm, the improved algorithm effectively reduces the time required for the analysis process and has good application value.**

Keywords—Water quality analysis; FCM clustering algorithm; Canopy-FFCM algorithm

I. INTRODUCTION

The ocean is a treasure trove of human resources[1]. The concentration of nutrients and other trace elements in seawater directly affects the growth, reproduction and metabolism of red tide organisms, and the deterioration of marine water environment affects the survival of aquatic organisms, which has become the focus of high attention of marine environmentalists for a long time[2]. How to quickly analyze the current water quality information is the premise of marine resource management and marine pollution control. With the improvement of information technology, the way of water environment monitoring has changed from manual collection and experimental analysis to automatic monitoring and analysis of water information. The automatic monitoring system of water environment has developed rapidly, but its ability to analyze data is still insufficient. Water quality information is stored in the monitoring system. These data are interrelated and have the characteristics of mass and redundancy. Analysis and modeling of these data are made in an attempt to obtain water quality pollution results from the data and predict its future development trend.

Fuzzy cluster analysis is a multivariate statistical method for quantitative study of classification problems[3]. The method determines the degree of each data point belonging to a certain cluster by using membership degree, which has been widely applied in the fields of image processing and pattern recognition, and is also a very important method in the field of water quality analysis. FCM algorithm is an improvement of HCM algorithm. It introduces a fuzzy concept to the attribution of each data and retains more detailed information. Its simple structure and fast computation speed have attracted the attention of many researchers. FFCM algorithm combines the characteristics of FCM algorithm and HCM algorithm, and uses HCM clustering algorithm with very fast operation speed to calculate the hard clustering center[4], which improves the efficiency of the overall algorithm[5].

ROV (Remote Operated Vehicle) is a kind of unmanned underwater vehicle, and it is the universal operating platform of underwater sensor network. In this paper, ROV is used to acquire real-time marine water quality data, as shown in Figure 1. Because the data collected by ROV data collection platform has the characteristics of large quantity, large vector dimension and more noise interference, many problems will occur when using traditional clustering algorithm to analyze massive water quality data information. For example, in the calculation process of Fuzzy C-means clustering algorithm, it is required that the sum of membership degrees of a data set is equal to 1, that is, formula 1 is satisfied[6].

$$\sum_{i=1}^{M} u_{ij} = 1, \quad \forall j = 1,2,\cdots\cdots,N \tag{1}$$

However, when there is noise in the data and the noise is far away from the center of each sample, originally his membership degree for each class is small. Normalization causes him to have a larger membership degree for each class, thus causing great errors and making iteration results unable to converge. When the improved FFCM algorithm is used to analyze the water quality data, although the iteration speed is accelerated, the initial optimal clustering number of the algorithm depends on manual regulation, and the stability of the algorithm is not high. This paper optimizes the initial iteration center of FCM algorithm and

the random selection of the optimal clustering number of FFCM algorithm, and proposes an improved algorithm, which can adapt to the characteristics of ocean water environment monitoring data and calculate the water quality more stably and quickly.

II. Related Algorithm

A. Traditional fuzzy c-mean clusering algorithm

Traditional fuzzy c-mean clustering algorithm uses membership degree to determine the degree to which each data point belongs to a certain cluster, which is a clustering algorithm with fuzzy partition. Fuzzy weighting index is an important parameter in FCM algorithm, and its value range is [1,+∞]. The data vectors （X_1， $X_2 \cdots X_n$）are divided into c fuzzy groups (A_1、 $A_2 \cdots A_c$), and each data point is determined to belong to each group by the membership degree between 0 and 1. The membership matrix U satisfies the normalization rule, and the sum of membership degrees of the data sets is 1, which satisfies Formula 2.

$$\sum_{i=1}^{a} u_{ij} = 1, \quad \forall j = 1,2,\cdots\cdots,\mathrm{n} \tag{2}$$

The algorithm steps are as follows:

1) Initialize membership matrix;
2) Calculate cluster centers A_i, $i = 1,\cdots,c$;
3) Calculate the value function;
4) update membership matrix.

B. Fast fuzzy c-mean clusering algorithm

In the process of using FCM algorithm, the values of membership matrix and clustering center are uncertain, and there is no fixed formula to get them. The value of the initial iteration center depends on human experience, and the performance of the algorithm depends on the initial clustering center. The calculation cannot ensure that FCM converges to an optimal solution. When facing a large number of data with high dimensions, we must use different initial clustering centers to run FCM algorithm many times, and the speed of data analysis cannot be guaranteed. Fast fuzzy c-mean(FFCM) clustering algorithm gives the selection scheme of initial iteration center, which effectively improves the problem of FCM algorithm. Its basic principle is: firstly, the hard clustering center of input data is determined by HCM algorithm[7], which is used as the initial iteration center of FCM algorithm, and FCM algorithm calculation is carried out. Finally, the convergence condition is satisfied, and the algorithm is finished. The steps are as follows:

1) Determine the number of clusters, Initialize membership matrix U_0, $u_{ij}\epsilon\{0,1\}$;
2) Calculate clustering center and modify membership matrix U_t;
3) If $|U_t - U_{t-1}| \leq \varepsilon$, The clustering center is taken as the initial iteration center of FCM algorithm and enters the iteration process of FCM algorithm. If not, repeat the second step.
4) update membership matrix.

For multi-dimensional water quality data, the hard clustering center calculated by HCM algorithm is used as FCM initialization data, which accelerates the iteration speed of FCM algorithm and improves the efficiency of water quality data analysis. However, the optimal clustering number still needs to be determined artificially, which depends on experience and has poor stability. Therefore, there is still a need for a method to improve FFCM algorithm.

C. Canopy clustering algorithm

With the development of traditional clustering algorithms to a certain period, it is difficult to meet the requirements in efficiency, and there is no scientific calculation formula for the initial clustering center. To solve the above problems, Andrew McCallum et al. proposed an improved clustering algorithm, namely Canopy algorithm[8]. Canopy algorithm can perform rough clustering on data in the preprocessing stage, and the implementation process is fast and simple[9]. The data is compared with two distance thresholds, and objects with similarity are placed in the same subset to obtain a series of overlapping canopy subsets. Its advantage lies in allowing overlap between subsets, increasing the fault tolerance of the algorithm and reducing isolated points in the results. The number of clusters does not need to be specified in advance, only the distance between the samples existing in the overlapping subset and the sample center point needs to be calculated, which can enhance the accuracy of clustering and reduce the calculation amount. It has great application value. The algorithm description diagram and algorithm effect diagram are as follows:

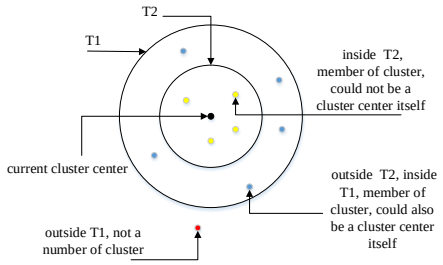

Fig. 1. Canopy algorithm description diagram.

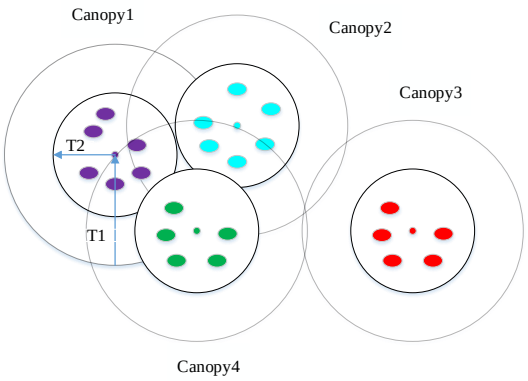

Fig. 2. Canopy algorithm effect diagram.

III. Algorithm Design Of Canopy-FFCM

A. Introduction of algorithm

The traditional FCM algorithm needs to set its own clustering center and number of clusters before it runs. Blind values will cause too many iterations of the algorithm and the

results cannot converge, which will greatly affect the efficiency of data analysis. Moreover, the convergence condition of the algorithm is that the value equation J(U, c1,c2…) is less than a certain threshold. The equation calculates the distance between the clustering centers and the data points without considering the influence of the clustering centers on each other. In the process of water quality data analysis, the sample data is large. If the value equation has multiple extreme points, when the iteration center reaches these extreme points, local convergence will occur and the calculation result will deviate from the expectation. Moreover, when the amount of data is large, FCM algorithm has high complexity, large amount of computation, and convergence speed cannot meet the requirements.

From the previous analysis, it can be seen that the Canopy algorithm does not need to determine the initial iteration center and clustering times when starting, but obtains the final result by calculating the similarity between the data. When the distance threshold of the algorithm is selected appropriately and the overlap rate between each Canopy class is not high, then the calculation amount for calculating the distance between the samples existing in the overlapping subset and the center point of the samples will be reduced, and the overall algorithm complexity will be reduced. In addition, a more accurate clustering number is obtained through the Canopy algorithm, thus avoiding the problem of randomly selecting the clustering number in the FCM algorithm. FFCM algorithm combines the advantages of fewer iterations of HCM algorithm and good clustering effect of FCM. Firstly, hard clustering centers are quickly obtained by HCM algorithm, and then fuzzy clustering is carried out by FCM, thus solving the problem of difficult selection of initial iteration centers of FCM algorithm. If Canopy algorithm is combined with FFCM algorithm, the whole algorithm will improve the accuracy of the results, improve the efficiency of the algorithm, and eliminate noise interference to some extent. Algorithm flow chart is as follows:

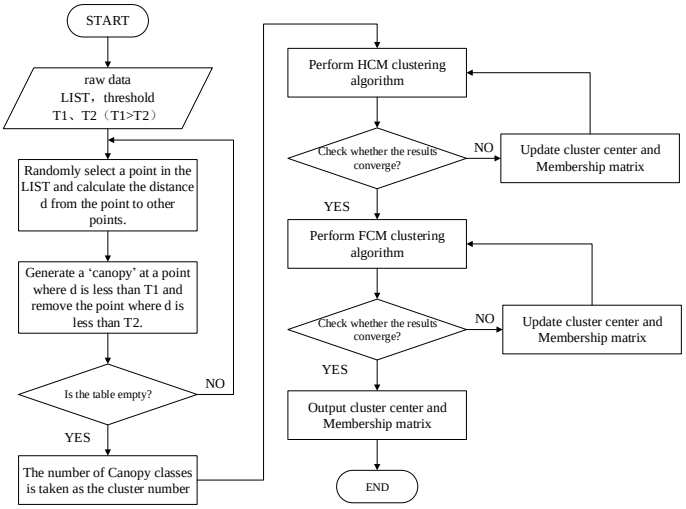

Fig. 3. Flow chart of Canopy-FFCM algorithm.

B. Rough Clustering Using Canopy Algorithm

In the initialization stage of Canopy algorithm, we don't need to determine the initial clustering center and cluster number. Moreover, the category is determined by calculating the distance between data, with less calculation. Before the algorithm is started, two distance thresholds T_1、T_2（$T_1 > T_2$）must be determined. T1 affects the number of classes and T2 affects the coverage of various classes. If T_1 and T_2 have reasonable values, the clustering effect is good. Through searching some literatures, the selection rules of $T_1$and T_2 in this paper are based on "minimum and maximum principle". Canopy-FCM algorithm flow for South China Sea water quality analysis is as follows:

1) Put the original data $D_i(i = 1 \cdots n)$ in an array named D. Set distance thresholds T_2 by formula 3 and T_1 is slightly larger than T_2.

$$T_2 = \frac{D_{max} - D_{min}}{2} \tag{3}$$

2) Take a random point as sample center M_1, and remove point M_1 from array D. Calculate the distance L_i ($i = 1 \cdots n$)from all points in array D to M_1.

3) The points X_i which satisfy formula 4 will be classified into the class named $Canopy_1$. Then M_1 is the cluster center of $Canopy_1$, and X_i is a member of $Canopy_1$.

$$0 < |X_i - M_1| < T_1,\ X_i \in V \tag{4}$$

4) Remove the points Y_k that satisfy formula 5 from array D. Then Y_k only exists in the $Canopy_1$ class and it cannot be the center point of other class.

$$0 < |Y_k - M_1| < T_2, Y_k \in Canopy_1 \tag{5}$$

5) Repeat Step 2 to Step 4 until the array D is empty[10]. Determine multiple clusters and the Canopy class number is c.

C. Initialize FFCM algorithm using Canopy class number

In this paper, Canopy algorithm is integrated into FFCM algorithm, and the center points of coarse clustering samples are calculated by the calculation process in the previous section. The number of center points is taken as the number of clusters initialized by FFCM algorithm. Initializing the FFCM parameter with the result of Canopy rough clustering not only solves the problem that FCM algorithm itself depends on the selection of initial iteration center and has high complexity, but also solves the problem that the number of FFCM clusters depends on human experience, and at the same time eliminates the influence of noise interference to some extent. The algorithm steps are as follows:

1) After the calculation in the previous section, the initial cluster number is equal to the constant c. Set the fuzzy weighting exponent m to 2 and initialize the membership matrix $U^{(0)}(u_{ij}\epsilon\{0,1\})$. Set iteration number t=0.

2) calculate the clustering center $A_i^{(t)}$ by using formula 6.

$$a_i^{(t)} = \frac{\sum_{j=1}^{n} {u_{ij}}^{(t)} x_j}{\sum_{j=1}^{n} {u_{ij}}^{(t)}}, i = 1, \ldots, c \tag{6}$$

3) Update Membership Matrix $U^{(t)}$ with Clustering Center.

$${u_{ij}}^{(t)} = \begin{cases} 1, \ \forall k \neq i, \|x_j - a_i\|^2 \leq \|x_j - a_k\|^2 \\ 0, \ \text{others} \end{cases} \tag{7}$$

4) Calculate the value of $||U^{(t)} - U^{(t-1)}||$. If it is less than the iteration stop threshold, reset the iteration number t=0, and use the above result as the initial iteration center $A_i^{(0)}$ of FCM algorithm.

5) Calculate the membership matrix $U_{ij}^{(h)}$ by using the value of the initial iteration center. d_{ij} represents the degree of difference between data point x_j and clustering center a_i, and euclidean distance is used in the calculation process.

$$u_{ij}^{(h)} = \frac{1}{\sum_{k=1}^{c} \frac{d_{ij}}{d_{kj}}^{\left(\frac{2}{m-1}\right)}} \tag{8}$$

6) Adjust iteration center $A_i^{(h)}$ by using formula 9.

$$a_i^{(h)} = \frac{\sum_{j=1}^{n} (u_{ij}^{(h)})^m x_j}{\sum_{j=1}^{n} (u_{ij}^{(h)})^m}, i = 1, \dots, c \tag{9}$$

7) Calculate the value function of FCM by using formula 10 and judge the iteration conditions. If the calculation result is less than a certain threshold, the algorithm stops, otherwise it returns to step 5.

$$J(U, A_1, A_2 \cdots) = \sum_{i=1}^{c} \sum_{j}^{n} \left(u_{ij}^{(h)}\right)^m d_{ij}^2 \tag{10}$$

According to the above steps, FFCM algorithm is initialized by using the number of clusters calculated by Canopy algorithm, the value of hard clustering center is obtained at high speed by HCM algorithm, and then the data is fuzzy clustered. This algorithm has good clustering effect, few iterations, fast convergence speed, and improves the efficiency of clustering analysis to a certain extent. The result of the algorithm obtains the membership matrix and determines the degree to which the data points belong to each class, of which the category with the greatest degree is the category to which the data belongs. Through clustering analysis of multidimensional data collected by ROV, water quality information can be quickly obtained.

IV. Experimental Result And Analysis

ROV and wired sensors were used to collect water samples from Dongpo Lake of Hainan University. The collected sample data includes 10 water quality indexes such as conductivity, DO, CODMN, NH3-N, pH, chlorophyll a, etc[10]. According to the objective requirements, DO, CODMN and NH3-N are finally selected as the experimental reference indicators. Because the data collected by wired sensors have the phenomenon of data error or loss, after data collection, these noises are preliminarily processed. In this section, FCM algorithm, FFCM algorithm and improved Canopy-FFCM algorithm are used to simulate and analyze the processed experimental samples.

Firstly, Canopy algorithm is used to roughly cluster the data. We use Matlab to simulate and analyze it. The clustering effect is shown in the figure. It can be seen from the figure that there are 3 Canopy clustering centers in each graph, forming 3 Canopy classes. This shows that all the data are divided into 3 classes, thus we get the initial cluster number, which is the same as the number of Canopy classes, that is, c=3.

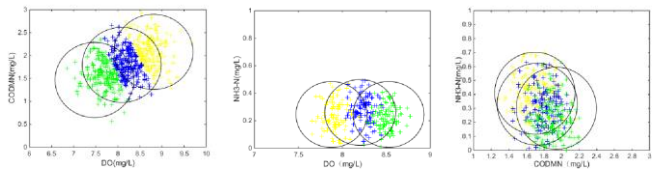

Fig. 4. Canopy clustering effect

By comparing the data calculated by HCM algorithm and FCM algorithm, we find that when the fuzzy weighting index is 2, the values of clustering centers obtained by the two algorithms are very close. Therefore, using HCM algorithm to calculate the clustering center of data first, and then applying the calculation results to the initial iteration center of FCM algorithm will solve the problem that FCM algorithm will consume more time under the condition of large amount of data. The following table shows the number of iterations run by the three clustering algorithms.

TABLE I. THE NUMBER OF ITERATIONS

Algorithm	Iteration number
FCM	69
HCM	12
Canopy-FFCM	28

Among them, HCM algorithm has the least number of iterations, which is 12. FCM algorithm requires 69 iterations for convergence process, while Canopy-FFCM algorithm requires 28 iterations, the first 12 of which are the number of iterations of HCM algorithm. The simulation results of Canopy-FFCM algorithm are shown in the following figure. It can be seen from the figure that the water quality of Dongpo Lake is most closely related to DO value, but is less affected by CODMN and NH3-N.

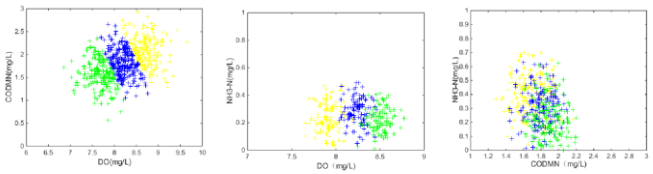

Fig. 5. Simulation results of Canopy-FFCM algorithm

Therefore, we find that when Canopy-FFCM is used to cluster data, the convergence time of the algorithm is much lower than that of FCM algorithm, which improves the operation efficiency of the overall model, especially in reducing the number of iterative operations, and the effect is particularly obvious.

V. Conclusion

As the quality of ocean water environment has gradually attracted the attention of world environmentalists, how to efficiently and rapidly analyze water quality information has become an important part of water environment monitoring. Aiming at the characteristics of high dimensionality and large data volume of marine water quality data, this paper combines the advantages of Canopy algorithm, HCM algorithm and FCM algorithm, and proposes Canopy-FFCM algorithm to improve the clustering number and initial iteration center. Experimental simulation proves the effectiveness of the algorithm in improving clustering efficiency. In this paper, the fuzzy weighting index still uses classical values. According to the characteristics of marine water quality, how to select a value more suitable for its characteristics needs further exploration.

Acknowledgment

This research received financial support from the Natural S cience Foundation of Hainan province (Grant #:617062), Natio nal Natural Science Foundation of China(Grant #: 61462022), Major Science and Technology Project of Hainan province (Gr ant #: ZDKJ2016015).

REFERENCES

[1] Milan Gocic,Slavisa Trajkovic, “Analysis of changes in meteorological variables using Mann-Kendall and Sen’s slope estimator statistical tests in Serbia”[J]. Global and Planetary Change. 2013

[2] Di Wu, Yu Zhang, Hao Wang, Mengxing Huang, Wenlong Feng, Rouru Chen, “Study on the assessment method of typhoon regional disaster based on the change of cholorophyll-a concentration in seawater,” OCEANS 2017 - Aberdeen, 2017

[3] Bezdek J C.Pattern, “recognition with fuzzy objective functional”[M], Plenum Press , New York , 1981

[4] K. A. Abdul Nazeer, M. P. Sebastian, “Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm”, In International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE -2009), Vol 1, July 2009, London, UK

[5] Hathaway R J, Bezdek J C , “Extending fuzzy and probabilistic clustering to very large data sets[J],” Computational Statistics & Data Analysis, 2006, 51(1):215-234.

[6] Tsai D M, Lin C C, “Fuzzy C-means based clustering for linearly and nonlinearly separable data”[J], Pattern Recognition, 2011, 44(8):1750-1760.

[7] Shehroz S.Khan, Amir Ahmad, “Cluster center initialization algorithm for K-modes clustering” [J]. Expert Systems with Applications, 2013, 40(18):7444-7456

[8] Mccallum A, Nigam K, Ungar L H, “Efficient clustering of high-dimensional data sets with application to reference matching” [J],Knowledge Discovery & Data Mining”, 2000:169-178.

[9] Zhang T,Ramakrishnan R,Livny M B, “An Efficient Data Clustering Method for Very Large Databases”, Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD 96) , 1996

[10] Sudipto Guha,Rajeev Rastogi,Kyuseok Shim, “CURE: An Efficient Clustering Algorithm for Large Databases”, Proceedings of the ACM SIGMOD International Conference on Management of Data, 1998