A Feature Preprocessing Framework of Remote Sensing Image for Marine Targets Recognition

Qiong CHEN
State Key Laboratory Marine Resource Utilization in South China Sea
College of Information Science and Technology
Haikou, China
13907534385@163.com

Mengxing HUANG
State Key Laboratory Marine Resource Utilization in South China Sea
College of Information Science and Technology
Haikou, China
huangmx09@163.com

Hao WANG
Big Data Lab, Department of ICT and Natural Sciences, Norwegian University of Science and Technology
Aalesund, Norway
hawa@ntnu.no

Yu ZHANG
State Key Laboratory Marine Resource Utilization in South China Sea
College of Information Science and Technology
Haikou, China
yuzhang_nwpu@163.com

Wenlong FENG
State Key Laboratory Marine Resource Utilization in South China Sea
College of Information Science and Technology
Haikou, China
fwlfwl@163.com

Xianpeng WANG
State Key Laboratory Marine Resource Utilization in South China Sea
College of Information Science and Technology
Haikou, China
wxpeng1986@126.com

Di WU
State Key Laboratory Marine Resource Utilization in South China Sea
College of Information Science and Technology
Haikou, China
hainuwudi@163.com

Uzair Aslam Bhatti
State Key Laboratory Marine Resource Utilization in South China Sea
College of Information Science and Technology
Haikou, China
uzairaslambhatti@hotmail.com

Abstract*-The effective extraction of continuous features of marine remote sensing image is the key to the processing of marine target recognition. Since many of the existing data mining algorithms can only deal with discrete attributes, it is necessary to convert continuous features into discrete features to adapt to these intelligent algorithms. In addition, most of the current discretization algorithms do not consider the mutual exclusion of attributes as well as that of breakpoints within an attribute when selecting breakpoints, and cannot guarantee the indistinguishable relation of decision table. So, it is not suitable for dealing with remote sensing data with multiple features obviously. Aiming at these problems, a feature preprocessing framework of remote sensing image for marine targets recognition is proposed in this paper. In the frame design of the whole preprocessing algorithm, the equivalent relationship model of information entropy is introduced to perform a series of comparison operations and loop controls, so as to obtain the optimal discretization interval number. Finally, simulation analysis is conducted on the high-resolution remote sensing image data collected in the port areas of South China Sea. The experiment shows that the framework proposed in this paper achieves excellent results in terms of interval number, accuracy, running time, and effectively detects the vessels targets at sea. Therefore, the proposed framework can be very well applied to the discretization of remote sensing image features for marine targets recognition.

Key words: Remote sensing image; data mining; feature preprocessing; information entropy; marine targets recognition

I. Introduction

Marine remote sensing refers to remote sensing regarding oceans and coastal zones as monitoring and research objects, including physical oceanographic remote sensing, biological oceanographic remote sensing and chemical oceanography remote sensing. Marine remote sensing uses a variety of sensors to make long-range, non-contact observations of the ocean to obtain images or data on oceanographic landscapes and oceanographic features[1]. To analyze and process the collected remote sensing image, it is necessary to firstly extract features which are continuous data in general. However, most of the current knowledge extraction methods and data mining algorithms can only deal with discrete data. Although some methods can handle continuous data, their performance is very poor. Therefore, it is necessary to convert continuous features into discrete features to adapt to these intelligent algorithms, reduce the time and space overhead of the algorithm, improve the system's ability to cluster samples, enhance the system's anti-noise ability, improve the algorithm's learning accuracy, and expand the application range[2].

However, most discretization algorithms still have relatively few applications in the analysis and processing of ocean optical remote sensing images, and all of them have certain defects[3], mainly in the following aspects: (1) the interdependencies between attributes are not taken into account when selecting breakpoints; (2) there are a large number of redundant breakpoints among attributes; (3) the lack of essential breakpoints leads to failure of the overall process; (4) do not consider the mutual exclusion of attributes as well as that of breakpoints within an attribute; (5) cannot guarantee the indistinguishable relation of decision table; (6) exponential growth of the program complexity, which is unable to meet the real-time dynamic marine target recognition processing. Obviously, none of these methods are suitable for processing remote sensing data with multiple features for marine targets recognition.

In order to solve the above problems, we proposed a feature preprocessing framework of remote sensing image for marine targets recognition in this paper. In the overall framework design of the preprocessing algorithm, the equivalent relationship model of information entropy was introduced. Firstly, we calculated the entropy for each interval. Secondly, each interval was scanned according to the set conditions, and the most suitable splitting position was found to generate a break point. Thirdly, the entire data preprocessing was controlled in loops by the evaluation of the results of before and after discretization, so as to obtain the optimal discretization interval number. Lastly, we conducted a simulation analysis of the high-resolution remote sensing image collected in the South China Sea port area. The experiment shows that the framework proposed in this paper achieves excellent results in terms of interval number, accuracy, running time, and effectively detects the vessels targets at sea. Therefore, the proposed framework can be very well applied to the discretization of remote sensing image features for marine targets recognition.

This paper is composed of four section. The remaining sections are organized as follow. Section 2 describes the basic concepts for the proposed work. Section 3 introduces the design structure of the entire feature preprocessing framework. The experimental results and discussion are presented in Section 4. Section 5 concludes this paper.

II. METHODOLOGY

A. Data Discretization

In simple terms, discretization is to adopt a specific method to divide a continuous interval into a limited number of cells, then to associate these cells with a set of discrete values. The discretization of continuous features (also called continuous attributes) is an important preprocessing step for data mining and machine learning, and is directly related to the effect of mining or learning[4]. The basic process of discretization is shown in Figure 1.

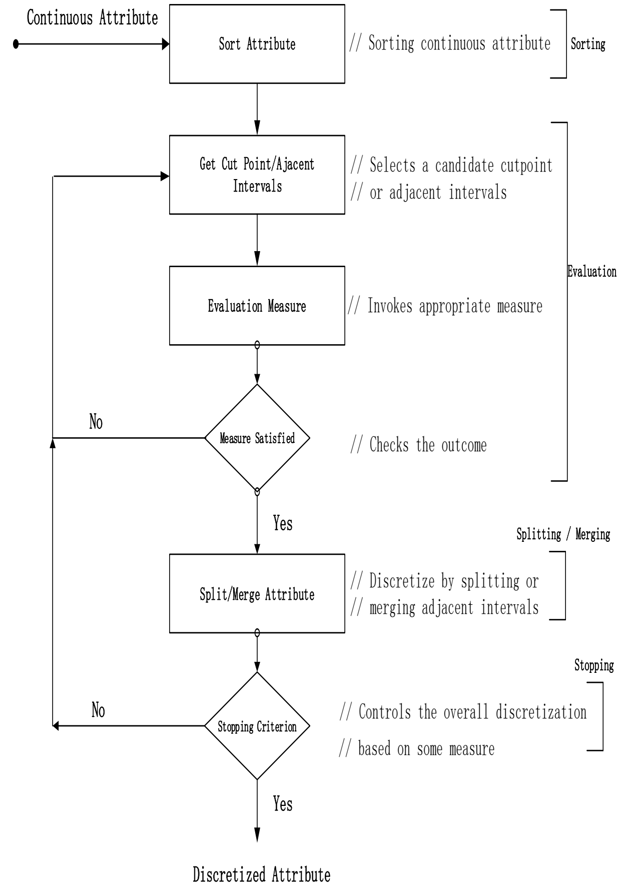

Fig. 1. Discretization process

First of all, the continuous attribute values to be processed (mainly referring to the pixel values in each band of the remote sensing image) are sorted according to a certain specified rule, such as insertion sort, bubble sort, selection sort, quick sort, heap sort, shell sort, etc. Then, initially determine the dividing points of the continuous attributes, that is, the selection of initial breakpoints. The next step is to split or merge breakpoints according to the discretization algorithm. Finally, if the termination condition is satisfied by the previous step, the whole discretization process is terminated, otherwise, returns to the previous step.

B. Rough Set Theory

Rough sets[5] are an important mathematical tool for handling uncertain data. In rough set theory, knowledge is regarded as the division of the universal, that is, knowledge is considered to be granular, and the uncertainty is caused by the large granularity in the knowledge. Different from the DS evidence theory[6] and the fuzzy set theory[7], the membership function value of the object in the rough set theory depends on the knowledge base. It can be directly obtained from the required data without any prior knowledge or additional information. So, it is much more objective to reflect the uncertainty of knowledge.

C. *Information Entropy Model*

Information entropy is a well-known mathematical theory proposed by Shannon, the father of information theory, for solving the quantitative measurement of information in the communication field[8]. Catlette, Fayyad, and Irani introduced information entropy into the discretization algorithm. According to the discussion of Fayyad and Irani, the formulas of information entropy and break point information entropy are given respectively.

$$E(S) = -\sum_{i=1}^{k} P(C_i, S)\log(P(C_i, S)) \quad (1)$$

$$E(A,T,S) = \frac{|S_1|}{|S|}Ent(S_1) + \frac{|S_2|}{|S|}Ent(S_2) \quad (2)$$

Where S is a set of objects, k is the number of categories, C_i represents the number of instances whose category is i in the set of objects S , A , T represents the breakpoint T on the attribute A , S_1 and S_2 represent the two objects sets of interval divided by breakpoint T respectively, $|S|$ denotes the cardinality of the set S .

III. REMOTED SENSING FEATURE PREPROCESSING FRAMEWORK

A. *The equivalent model of information entropy*

According to the theory of the previous chapter, we can establish an equivalent model of information entropy in the analysis and processing or remote sensing images. We suppose that S denotes a set of image pixels, k denotes the number of land cover categories, C_i denotes the number of instances of the category i in the pixel set S , A , T represent the break point T in the band A , S_1 and S_2 represent the two pixel sets of interval divided by the break point T in the band A respectively, $|S|$ represents the cardinality of the set S , that is, the total number of pixels included in S . Then, in the information decision table, we suppose that U denotes the collection of image pixels, the attributes in condition attribute set C represent bands, D contains only one decision attribute that corresponds to the land cover class in the remote sensing image, V_a represents the value domain of the a th band. The equivalence class about attribute subset A in the universal U is also defined as the following formula.

$$U \mid IND(A) = \{X \mid X \subseteq U \wedge (\forall x \in X \, \forall y \in X \Rightarrow \forall a \in A(a(x) = a(y)))\} \quad (3)$$

B. *Remote sensing feature preprocessing framework*

According to the established equivalent model of information entropy, we designed a remote sensing feature preprocessing framework for marine targets recognition. As shown in Figure 2.

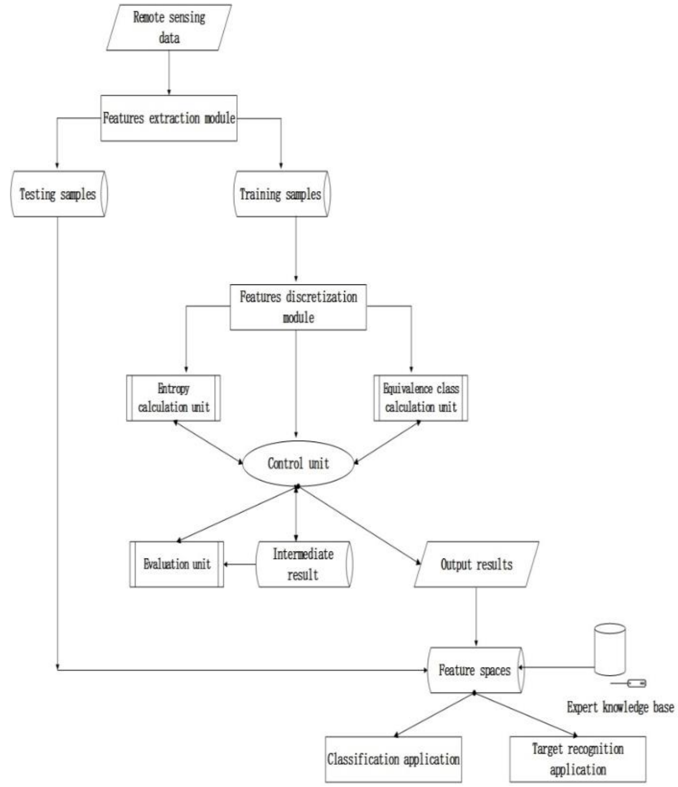

Fig. 2. Design structure of remote sensing feature preprocessing framework

The entire remote sensing image feature preprocessing framework consists of several functional modules, processing units and memories. At the beginning, remote sensing data is input into the system, and the feature extraction module reads out the data features and generates training samples and testing samples. Then, the training samples are sent to the feature discretization module for preprocessing. The feature discretization module consists of an entropy calculation unit, an equivalent class calculation unit, an evaluation unit, a control unit, and a memory for storing intermediate results. After the continuous data is discretized, the results are output. Together with the testing samples and domain knowledge provided by the expert knowledge base, the feature spaces of the remote sensing image are formed. In this feature space, the system can carry out the further classification applications and targets recognition applications.

IV. EXPERIMENTS AND ANALYSIS

A. *Data Source*

The experimental data used in this paper comes from a GF-2 satellite image in the offshore port area of Zhanjiang City, China, on October 7, 2015, which contains four bands. The objects in this image are divided into six categories: ship, port, building, bare land shoal, water body and vegetation.

B. *Experimental Environment*

This experiment was performed on a computer with Intel(R) Core(TM) i5-5200U CPU@2.20GHz processor and 12G RAM hardware. Visualization, programming, simulation, testing and numerical calculation processing of this experiment are implemented in MATLAB (R2016a version) environment. Radiometric calibration of images, atmospheric correction, and comparison of results before and after discretization are performed under ENVI 5.3 environment.

C. Experimental Process and Results Analysis

Firstly, several regions covering six major categories are randomly selected from the image and integrated as training samples to be discretized, containing a total of 2150 pixels, among which 500 are ships, 517 are ports, 131 are buildings, 99 are bare land shoals, 804 are water bodies, 99 are vegetation. Then, after the pixels are sorted, and eliminates the duplicates by value within the band, the number of initial breakpoints for the four bands is obtained, which is 479, 476, 352, 335, respectively. Therefore, the training sample has a total of 1642 breakpoints at the beginning. We used the proposed method to discretize the above data. The results of the number of intervals for each band, equivalence class differences, data inconsistency, and system runtime are shown in Table 1 and Table 2.

Table. 1. Number of intervals

Band1	Band2	Band3	Band4	Total Intervals
374	374	336	284	1368

Table. 2. Method performance

Equivalence Class Differences	Data Inconsistencies	Program Running Time
0	3	41.8s

As shown in Table 1 and Table 2, we can see that after the samples was discretized under the framework of remote sensing feature preprocessing in this paper, 1368 breakpoints were obtained, 274 fewer than the original number, the number of equivalence class differences was zero, and the number of errors was only 3, the speed was ideal.

The result of the classification accuracy evaluation at the pixel level is usually represented by the confusion matrix. As well as confusion matrix, Kappa coefficient[9] is also widely used in remote sensing image classification accuracy evaluation. Based on the confusion matrix, it quantifies the overall effectiveness of the classifier. Table 3 shows the evaluation indicators produced by the analysis of classifier for different equivalence class differences under the preprocessing framework of this article.

Table. 3. Classification accuracy under different number of differences

Differences	Accuracy	Kappa coefficient
0	85.2%	0.7873
2	83.3%	0.7675
4	74.2%	0.7060
6	70.6%	0.6760
8	65.7%	0.6362
10	60.2%	0.5556
12	56.5%	0.5286

As can be seen from the Table 3, the accuracy of classification decreases as the number of equivalence class differences between before and after discretization increases. Figure 3 shows the accuracy of different number of equivalence class differences when the band is discretized.

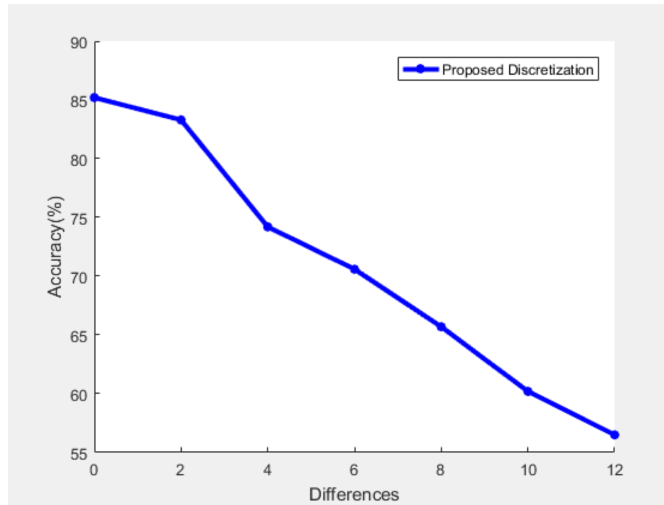

Fig. 3. Relationship between differences and accuracy

The Figure 3 shows that the degree of change in the number of equivalence class differences in the decision table has a large impact on the accuracy of the classification. We have also achieved good results in the use of the discretization under the remote sensing feature preprocessing framework for vessel recognition in image[10]. Figure 4 is a partial view of the image. It shows that the docking vessel can be effectively separated at the boundary of the port.

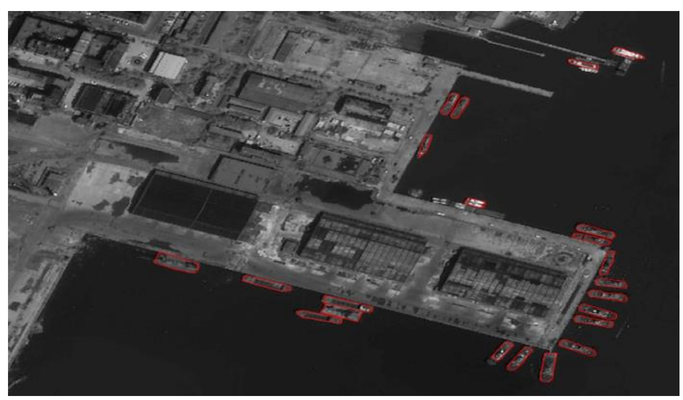

Fig. 4. Result of marine vessel targets recognition

V. CONCLUSIONS AND FUTURE WORK

In this paper, a feature preprocessing framework of remote sensing image for the marine targets recognition is proposed to solve the problem of discretization of marine remote sensing data with multiple features. Simulation experiments verify the effectiveness of the proposed method. It provides a new idea for preprocessing of remote sensing image. It also brings certain guiding significance to the analysis and design of the discretization methods in the marine targets recognition application.

Future research work includes: (1) Compare this method with other discretization algorithms to improve the entire technical framework of feature preprocessing; (2) Apply the framework of this article to other data sets for further testing and improvement to make it more practical.

ACKNOWLEDGMENTS

Foundation item: National Natural Science Foundation of China(Grant #: 61462022), the National Key Technology Support Program (Grant #: 2015BAH55F04, Grant #:2015BAH55F01), Major Science and Technology Project of Hainan province (Grant #: ZDKJ2016015), Natural Science Foundation of Hainan province (Grant #: 20156235, Grant #: 614232 and Grant#:617062), Higher Education Reform Key Project of Hainan province (Hnjg2017ZD-1), Scientific Research Staring Foundation of Hainan University(Grant #: kyqd1610)..

REFERENCES

[1] Wenkai Li, Kun Fu, Hao Sun, Xian Sun, Zhi Guo, Menglong Yan, Xinwei Zheng. Integrated Localization and Recognition for Inshore Ships in Large Scene Remote Sensing Images[J]. IEEE Geoscience and Remote Sensing Letters, 2017,14(6):936-940.

[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 [C]//OCEANS. Aberdeen, UK: IEEE, 2017.

[3] Salvador Garcı ´ a, Julia ´ n Luengo, Jose ´ Antonio Sa ´ ez, Victoria Lo ´ pez, Francisco Herrera. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning[J]. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2013,25(4):734-750.

[4] Xu He, Fan Min, William Zhu. Comparison of Discretization Approaches for Granular Association Rule Mining[J]. Canadian Journal of Electrical and Computer Engineering, 2014,37(3):157-167.

[5] Zdzislaw Pawlak. Rough Set: Theoretical Aspects of Reasoning about Data[M]. USA: Kluwer Academic Publishers, 1992.

[6] Sun Yanhua, Yang Yi, Li Yonghua, Zhang Yanhua. Full diversity reception based on Dempster-Shafer theory for network coding with multiple-antennas relay[J]. China Communications, 2015,12(10):76-90.

[7] G.A. Vijayalakshmi Pai. Fuzzy Decision Theory Based Metaheuristic Portfolio Optimization and Active Rebalancing Using Interval Type-2 Fuzzy Sets[J]. IEEE Transactions on Fuzzy Systems, 2017,25(2):377-391.

[8] Claude E Shannon. A mathematical theory of communication[J]. Bell System Technical Journal, 1948,27(3):379-423.

[9] Susana M. Vieira, Uzay Kaymak, João M. C. Sousa. Cohen’s kappa coefficient as a performance measure for feature selection[C]//Fuzzy Systems (FUZZ). Barcelona, Spain: IEEE, 2010.

[10] Vikas Patel, H. Madhukar, Sujatha Ravichandran. Variability index constant false alarm rate for marine target detection[C]//Signal Processing And Communication Engineering Systems (SPACES). Vijayawada, India: IEEE, 2018.