Chen and Huang *Journal of Cloud Computing: Advances, Systems and Applications* (2021) 10:5
https://doi.org/10.1186/s13677-020-00216-4

Journal of Cloud Computing: Advances, Systems and Applications

RESEARCH Open Access

Rough fuzzy model based feature discretization in intelligent data preprocess

Qiong Chen[1,2*] and Mengxing Huang[1,2*]

Abstract

Feature discretization is an important preprocessing technology for massive data in industrial control. It improves the efficiency of edge-cloud computing by transforming continuous features into discrete ones, so as to meet the requirements of high-quality cloud services. Compared with other discretization methods, the discretization based on rough set has achieved good results in many applications because it can make full use of the known knowledge base without any prior information. However, the equivalence class of rough set is an ordinary set, which is difficult to describe the fuzzy components in the data, and the accuracy is low in some complex data types in big data environment. Therefore, we propose a rough fuzzy model based discretization algorithm (RFMD). Firstly, we use fuzzy *c*-means clustering to get the membership of each sample to each category. Then, we fuzzify the equivalence class of rough set by the obtained membership, and establish the fitness function of genetic algorithm based on rough fuzzy model to select the optimal discrete breakpoints on the continuous features. Finally, we compare the proposed method with the discretization algorithm based on rough set, the discretization algorithm based on information entropy, and the discretization algorithm based on chi-square test on remote sensing datasets. The experimental results verify the effectiveness of our method.

Keywords: Feature discretization, Preprocessing technology, Edge-cloud computing, Fuzzy *c*-means, Rough fuzzy model

Introduction

Edge-cloud computing is based on the core of cloud computing and the capability of edge computing, forming an elastic cloud platform built on the edge infrastructure [1–3]. As an extension of the centralized cloud, the edge cloud provides low-latency, self-organizing, and schedulable distributed cloud services for terminals [4, 5]. As shown in Fig. 1, the edge cloud, the centralized cloud and the terminals of Internet of things constitute an end-to-end technical architecture of "cloud-edge-terminal collaboration". By allocating computing, network forwarding, storage, and other work to the edges for intelligent data preprocessing, the cloud pressure, response delay and bandwidth cost can be reduced [6, 7]. Feature discretization is an important reduction technology for mass data in industrial control [8, 9]. It can filter abnormal data, reduce system load, and improve the performance of intelligent algorithm [10] by transforming continuous features into discrete ones that are easier to understand, use, and interpret, so as to improve the efficiency of edge-cloud computing and prevent network attacks to a certain extent [11, 12].

In recent years, feature discretization has gradually become a key technology of intelligent data preprocessing, which has attracted extensive attention all over the world and achieved fruitful research results [13]. Obtaining the optimal discretization scheme has been proved to be an NP complete problem [14]. Most of the current methods are based on specific partition criteria to realize the discretization of continuous features, such as the equal width algorithm [15], the equal frequency algorithm [15], the discretization algorithm based on information entropy [16], the discretization algorithm based

* Correspondence: 13907534385@163.com; huangmx09@163.com
[1]State Key Laboratory of Marine Resource Utilization in South China Sea, Haikou 570228, China
Full list of author information is available at the end of the article

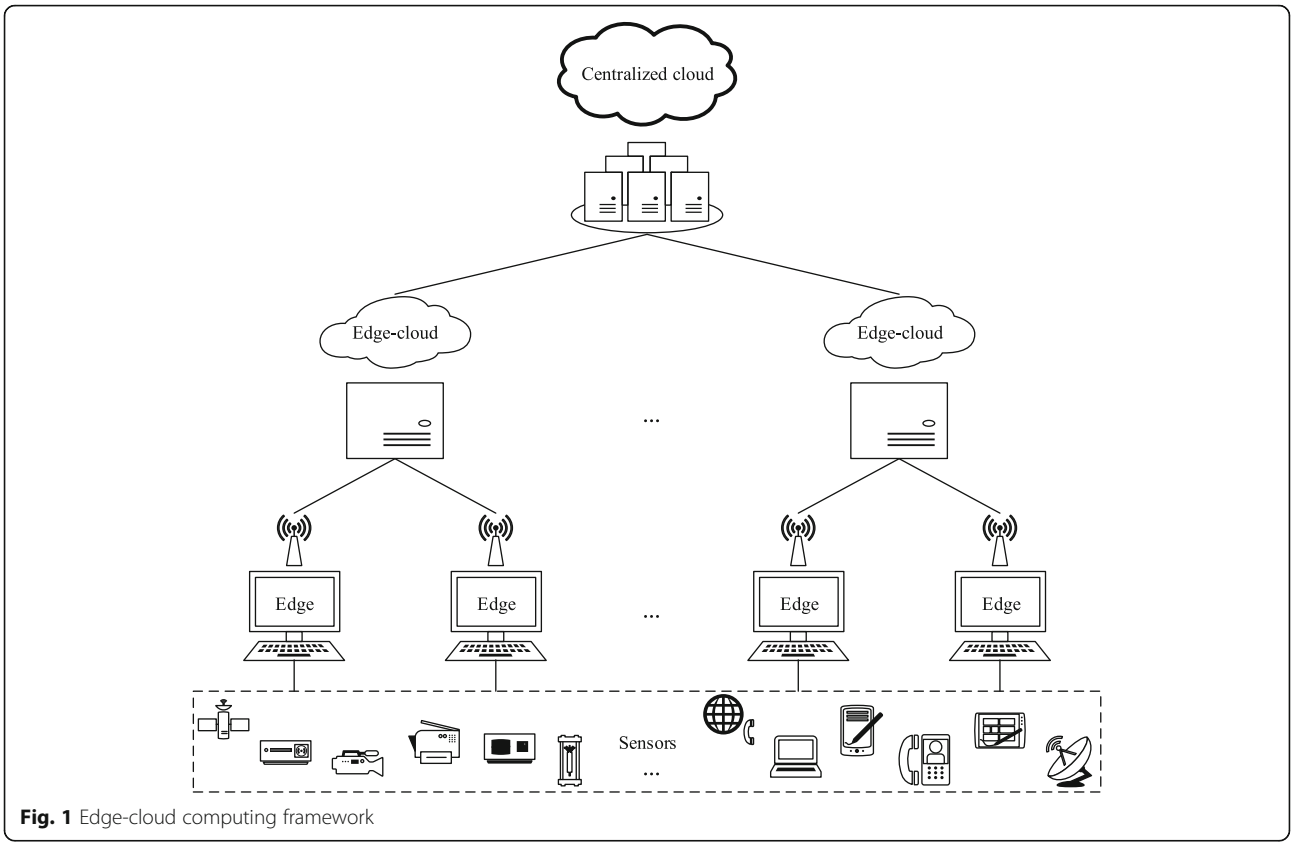


Fig. 1 Edge-cloud computing framework

on chi-square test [17]. However, due to the complex correlation between features, the relatively fixed partition criteria cannot comprehensively measure the discrete interval. In addition, the distribution of sample attribute values in a dataset is often difficult to learn. Therefore, the discretization results obtained by these algorithms are often not the optimal scheme in specific application scenarios, and even fail to meet the accuracy requirements of the system [18].

Compared with the above discretization methods, discretization based on rough set [19] has achieved good results in many applications because it can make full use of the known knowledge base without any prior information. On the other hand, since feature discretization is a complex constrained optimization problem [13], it is very difficult to solve this kind of problem by traditional methods, and the genetic algorithm is more effective than traditional methods because of its group search strategy and calculation method that is not dependent on gradient information [20]. Through crossover and mutation operations, genetic algorithm takes into account the global and local equilibrium search ability. Compared with other swarm intelligence optimization algorithms, genetic algorithm can use more mature analysis methods to estimate the convergence rate [21]. Therefore, the combination of rough set and genetic algorithm can obtain better results than other methods. Chen et al. propose a genetic algorithm for discretization [22]. They conduct experiments on several datasets in UCI machine learning library. In the experimental process, they use some optimization strategies to continuously optimize the genetic algorithm. The experimental results show that the genetic algorithm is effective in both time complexity and accuracy. Ren et al. propose a heuristic genetic algorithm to discretize continuous attributes of decision table [23]. The algorithm takes the importance of continuous cut sets as heuristic information, and constructs a new operator, which not only keeps the identifiability of the selected cut sets, but also improves the local search ability of the algorithm. Dai uses the rough set model to construct the individual fitness function of genetic algorithm to evaluate the uncertainty of information system, so as to handle the consistency and minimum [24]. With the advantage of rough set in dealing with incomplete information, the above methods can use the strong search ability of genetic algorithm to obtain the minimum number of breakpoints while ensuring that the compatibility of the system is not destroyed. However, in big data environment, there are often a large number of complex types of data, and the uncertainty in decision-making is caused by the unclear classification of categories. The equivalence class of rough set is an ordinary set, which is difficult to describe the

fuzzy components in the data, and the accuracy obtained in these complex data types is low. Fuzzy set is a mathematical tool used to describe fuzziness, and the combination of fuzzy set and rough set can better deal with the uncertainty of data [25].

For this reason, we propose a rough fuzzy model based discretization algorithm (RFMD). The main contributions of this article are as follows: (1) we create a fuzzy set for each category in the dataset, and use fuzzy *c*-means [26] to get the membership function of each category; (2) we use the membership function to fuzzify the equivalence relationship of rough set, and establish the fitness function of genetic algorithm [18] based on rough fuzzy model [27] to select the best breakpoints on continuous features.

The rest of this paper is arranged as follows: the second part introduces the basic concepts of feature discretization, rough set, and fuzzy set; the third part describes the discretization algorithm based on rough fuzzy model; the fourth part introduces the experimental environment and datasets, and analyzes and discusses the experimental results; the fifth part summarizes the full text.

Background

We inntroduce the basic process of feature discretization, and the binary coding of feature discretization in genetic algorithm. Then, we explain the related definitions of rough sets and fuzzy sets, and lead to the rough fuzzy model.

Feature discretization and genetic coding

Discretization is to divide the continuous features (also known as continuous attributes) into a finite number of subintervals by some specific method, and associate these subintervals with a group of discrete values (also known as breakpoints) [28]. Through discretization, the data scale can be greatly reduced, thus improving the efficiency of massive data processing at the edge nodes of edge-cloud computing, and greatly relieving the pressure of transmitting data back to the centralized cloud [11]. The basic process of feature discretization is shown in Fig. 2.

In the beginning, the values of continuous attributes are sorted and the duplicate values are deleted to get a set of candidate breakpoints; then, the partition breakpoints of the continuous attributes are selected from the candidate breakpoints set, and decide whether to divide the interval or merge adjacent subintervals according to the judgment criteria of the discretization algorithm; if the termination condition is satisfied, the discretization result is output; otherwise, the remaining breakpoints are selected from the candidate breakpoints set to perform discretization of attributes.

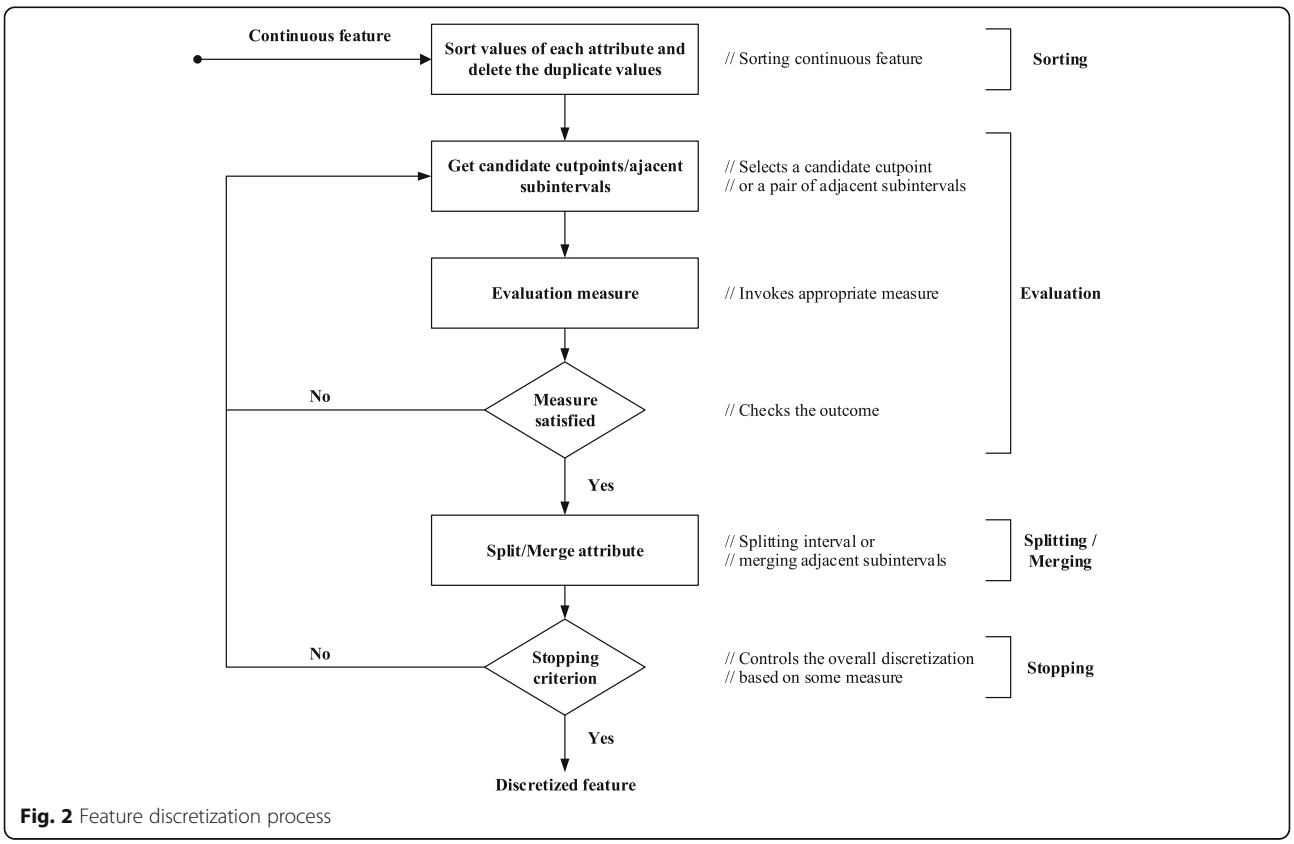


Fig. 2 Feature discretization process

Genetic algorithm is a probabilistic evolutionary algorithm for global optimization [29], which has achieved good performance in many optimization problems [30]. Genetic algorithms use fitness function to evaluate the quality of individuals in population, and transform the problem-solving process into a process which is similar to the crossover and mutation of chromosomal genes in biological evolution. In many complex combinatorial optimization problems, genetic algorithm can quickly obtain better optimization result than some conventional optimization algorithms [20, 21]. However, genetic algorithm cannot directly deal with the parameters of the problem space, so the problem to be solved must be expressed as a chromosome or individual in genetic space by coding. This conversion is called genetic coding [30]. Genetic coding adopts the following criteria [18]: (1) completeness: all candidate solutions in the problem space can be represented as chromosomes in genetic space; (2) soundness: chromosomes in genetic space can correspond to all candidate solutions in the problem space; (3) non-redundancy: chromosomes and candidate solutions are one-to-one correspondence.

The discretization problem can be seen as the selection of candidate breakpoints [30]. Each chromosome in the population represents a possible discretization scheme. The length of chromosome is equal to the number of candidate breakpoints. We use binary coding to encode the candidate breakpoints. Each bit in the binary code corresponds to a candidate breakpoint. The values of '1' and '0' represent that the corresponding breakpoint is selected and not selected, respectively. The set of selected candidate breakpoints is a possible discretization scheme.

Rough sets

Rough set is a mathematical theory proposed by Pawlak to solve the problem of data uncertainty [31]. Rough set regards knowledge as the ability to classify objects in the universe. An equivalence relation on the universe represents a knowledge.

Definition 2.1

The two-tuple $K = (U, \mathbb{R})$ is a knowledge base, where, U is the universe, and $\mathbb{R}$ is the cluster of equivalence relations on U.

Definition 2.2

For $x \in U$, $R \in \mathbb{R}$, the equivalent class of x under R is $[x]_R = \{y \in U | (x, y) \in R\}$. The quotient set $U/R = \{[x]_R | x \in U\}$ is called a knowledge.

Definition 2.3

Suppose U is a non-empty finite universe, and R is a binary equivalence relation on U. For any $X \subseteq U$, the lower and upper approximations of X with respect to R are:

$$R\_X = \{x \in U | [x]_R \subseteq X\} \tag{1}$$

$$R^{-}X = \{x \in U | [x]_R \cap X \neq \varnothing\} \tag{2}$$

Discretization based on rough set evaluates the result of discretization according to the degree of dependence of X on R. The degree of dependence of X on R is:

$$\gamma_R(X) = \frac{|R\_X|}{|U|} \tag{3}$$

Where, $|\cdot|$ is the cardinality of the set. It is easy to see that discretization based on rough set can make full use of the known knowledge base without any prior information. However, $[x]_R$ is an ordinary set, which is difficult to describe the fuzzy components in data.

Rough fuzzy-model

Fuzzy set is a mathematical theory proposed by Zadeh to describe the fuzziness of data [32]. Compared with the ordinary set which can only express crisp concepts, fuzzy sets can represent not only crisp concepts, but also fuzzy concepts.

Definition 2.4

Let A be a mapping from set X to $[0, 1]$, call A the fuzzy set on X, and function $A(x)$ is the membership of x to the fuzzy set A. The fuzzy set A is expressed as follows when X is a finite set and when X is an infinite set:

$$A = \sum_{i=1}^{n} A(x_i)/x_i \tag{4}$$

$$A = \int_X A(x)/x \tag{5}$$

Through the membership function $A(x)$, the equivalent classes of rough sets can be fuzzified to obtain the rough fuzzy model [33]. If X is a finite set, then the cardinality of fuzzy set A is:

$$|A| = \sum_{x \in X} A(x) \tag{6}$$

Definition 2.5

Let U be the non-empty finite universe, R is the binary equivalence relation on U, and A is the fuzzy set on U. For any $x \in U$, the lower and upper approximations of x in the rough fuzzy model established by R and A are:

$$R\_A(x) = \inf_{y \in U} \{A(y) | (x, y) \in R\} \tag{7}$$

$$R^{-}A(x) = \sup_{y \in U}\{A(y) | (x,y) \in R\} \tag{8}$$

Accordingly, the approximate accuracy of the above rough fuzzy model is:

$$\eta = \frac{|R\_A|}{|R^{-}A|} \tag{9}$$

Since $R\_A(x) \le R^{-}A(x)$, $0 \le \eta \le 1$. The closer the value of η to 1, the higher the overall approximation accuracy. In the application process of edge-cloud computing, the massive data collected often have incomplete, fuzzy, and other uncertain information. Rough fuzzy model has the advantages of both rough set and fuzzy set. It can make full use of the known knowledge base without any prior information, and use membership function to fuzzify the equivalent relationship to describe the fuzzy components inside the data, so as to improve the accuracy of the massive data processing at the edge node of edge-cloud computing [27, 34].

Rough fuzzy model based discretization algorithm

We introduce the process of calculating membership by fuzzy *c*-means clustering. Then, we detail the fitness function based on rough fuzzy model. Finally, we describe the whole process of the proposed method.

Membership calculated by fuzzy *c*-means clustering

Fuzzy *c*-means integrates the essence of fuzzy theory [35]. Compared with the hard clustering of *k*-means, fuzzy *c*-means provides more flexible clustering results [36]. In most cases, the objects in the dataset cannot be divided into crisp clusters. It is hard to assign an object to a specific cluster, and errors may occur. Therefore, it is necessary to assign a weight between each object and each cluster to indicate the degree to which the object belongs to the cluster. Certainly, probability-based methods can also give such weights. But it is difficult for us to determine an appropriate statistical model. Therefore, it is a better choice to use the fuzzy *c*-means with natural and non-probabilistic characteristics [37].

The dataset can be represented by information table $S = (U, R, V, f)$. Where, U is the non-empty finite universe, R is the set of the attributes, V is the range of attribute values, and f is the mapping function from the object to the range of attribute values. Suppose that U contains N samples, C categories, M attributes, x_{ih} is the value of sample x_i on the h-th attribute, $1 \le i \le N$, $1 \le h \le M$, and the class center c_j of the j-th class is initialized to c_j^0, $1 \le j \le C$, then the membership of x_i to the j-th class is initialized as follows:

$$u_{ij}^0 = 1 \Big/ \sum_{k=1}^{C} \left(\frac{\sum_{h=1}^{M} (x_{ih} - c_{jh})^2}{\sum_{h=1}^{M} (x_{ih} - c_{kh})^2} \right) \tag{10}$$

Where, c_{jh} is the value of the current class center c_j on the h-th attribute. After the current membership is obtained, the class center c_j is updated to:

$$c_j^1 = \sum_{i=1}^{N} \left(\left(u_{ij}^0\right)^2 \times x_i \right) \Big/ \sum_{i=1}^{N} \left(u_{ij}^0\right)^2 \tag{11}$$

u_{ij} and c_j are updated iteratively until the following termination condition is met:

$$\max_{ij} \left\{ |u_{ij}^{t+1} - u_{ij}^{t}| \right\} < \varepsilon \tag{12}$$

Where, t is the number of iterations, and ε is the error threshold. In this way, the membership of each sample in U is obtained, as shown in Algorithm 1.

Algorithm 1: Calculation of membership

Input: Information table $S = \{U, R, V, f\}$, cluster centers $C = [c_j]$, error threshold ε

Output: Membership matrix $F = [u_{ij}]$

1: *Initialize* cluster centers and membership matrix;
2: **repeat**
3: $t = t + 1$;
4: Update $F^{t+1} = [u_{ij}^{t+1}]$ and $C^{t+1} = [c_j^{t+1}]$ by (10) and (11);
5: **until** $\max_{ij} \{|u_{ij}^{t+1} - u_{ij}^{t}|\} < \varepsilon$

Fitness function based on rough fuzzy model

After obtaining the membership of each sample in U to each category, we create a fuzzy set for each category:

$$A_j(x_i) = u_{ij}, 1 \le i \le N, 1 \le j \le C \tag{13}$$

Where, A_j is the corresponding fuzzy set of the j–th class. According to (7) and (8), we can calculate the lower and upper approximations of x_i in the rough fuzzy model established by attribute set R and A_j:

$$R\_A_j(x_i) = \inf_{y \in U}\{A_j(y) | (x_i, y) \in R\} \tag{14}$$

$$R^{-}A_j(x_i) = \sup_{y \in U}\{A_j(y) | (x_i, y) \in R\} \tag{15}$$

Accordingly, the average approximation accuracy of the rough fuzzy sets of all classes is:

$$\bar{\eta} = \frac{1}{C} \sum_{j=1}^{C} \frac{|R\_A_j|}{|R^{-}A_j|} \tag{16}$$

Since the optimal discretization scheme is the best trade-off between data consistency and the number of breakpoints [38]. Therefore, the fitness function should be determined by the average approximation accuracy and the number of breakpoints. Assuming that $|D|$ is the number of breakpoints reduced by discretization scheme D, the fitness function is as follows:

$$Fit = \alpha \times |D| + \beta \times \bar{\eta} \quad where\ \alpha \geq 0, \beta \geq 0, and\ \alpha + \beta = 1 \tag{17}$$

Where, α and β are weight coefficients. The selection of the mentioned parameters is an open problem, as no specific selection can adapt to all datasets. Generally, the rationality of parameters is judged according to the characteristics of datasets and experimental observation [39]. $|D|$ determines the magnitude of the reduction in the number of breakpoints, while $\bar{\eta}$ controls the accuracy of data. If α is much greater than β, the accuracy of data will be very low. If α is far less than β, the number of breakpoints will be large, so the purpose of discretization cannot be achieved. Generally, in order to obtain as few breakpoints as possible while ensuring the accuracy of data, $0.1 \leq \alpha \leq 0.5$, i.e., $0.5 \leq \beta \leq 0.9$. The purpose of this paper is to improve classification accuracy after discretization, and classification accuracy is directly related to the average approximation accuracy of rough fuzzy sets. Therefore, we set β to be larger than α ($\alpha = 0.1$, $\beta = 0.9$), and achieve good results in the experiment.

Based on this fitness function, we iteratively perform genetic operation to find the optimal breakpoint set on continuous features. The whole process is shown in Algorithm 2. At first the membership function of each category is obtained through Algorithm 1, and the fitness function based on rough fuzzy model is established. Then, for each individual in the population, the average approximation accuracy of the corresponding discretization scheme is obtained by calculating the upper and lower approximations of all samples. Finally, in each genetic operation, the fitness of all individuals in the population is calculated by the number of breakpoints and the average approximation accuracy of the discretization scheme, and the global variable is updated by the individual with the highest fitness. When the accuracy requirement of the system is met or the set number of iterations is exceeded, the program is stopped and the optimal discretization scheme is output. Otherwise, the genetic algorithm will continue to be executed until the termination conditions are met.

Algorithm 2: Rough fuzzy model based discretization

Input: Information table $S = \{U, R, V, f\}$, fuzzy sets $A = [A_i]$, limited iterations T, precision threshold δ

Output: Optimal discretization scheme D

1: $t = 0$;
2: *Initialize* population P(t) and global variable G;
3: **repeat**
4: Perform genetic operation on P(t) and update G with the individual with maximum fitness by (17);
5: $t = t + 1$;
6: **until** $(t > T) \vee (Fit(G) \geq \delta)$
7: $D = G$;

Rough fuzzy model versus rough model

Fuzzification and rough set in RFMD enable reasoning uncertainty problems. Figure 3 is a simple example to illustrate the advantages of RFMD over the discretization methods based on rough set. Suppose that the dataset contains three samples (x_1, x_2, and x_3), the corresponding attribute values are v_1, v_2 and v_3, and the corresponding categories are C_1, C_1, and C_2. Through fuzzy c-means, the membership degree of three samples to C_1 and C_2 are: $C_1(x_1) = 0.8$, $C_1(x_2) = 0.5$, $C_1(x_3) = 0.4$, $C_2(x_1) = 0.2$, $C_2(x_2) = 0.5$, $C_2(x_3) = 0.6$. When the dataset needs to be divided into two intervals, there are two discretization schemes to choose from, shown in Fig. 3b and c. We can see that the membership of x_2 is quite different from that of x_1. In comparison, the membership of x_2 is closer to that of x_3. Obviously, the division in Fig. 3c looks more reasonable.

We use RFMD and rough set-based discretization methods to discretize the original information table in Fig. 3a, and verify the effectiveness of RFMD by comparing the discretization results:

(1) RFMD selects the best discretization scheme by comparing the average approximation accuracy η_1 and η_2 of Fig. 3b and c. In Fig. 3b, the equivalence classes under *Attribute* are $\{x_1, x_2\}$ and $\{x_3\}$. According to (14)

Sample	Attribute	Class	Membership	
			C1	C2
x1	v1	C1	0.8	0.2
x2	v2	C1	0.5	0.5
x3	v3	C2	0.4	0.6

(a) Original information table

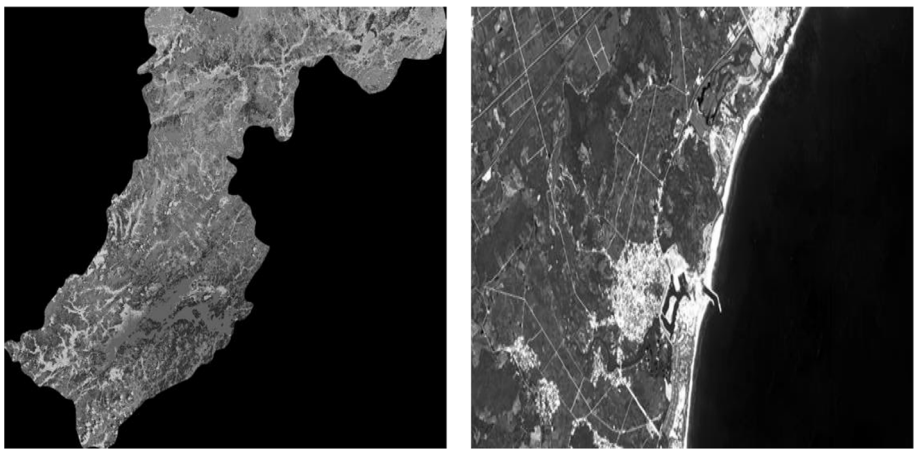
Sample	Attribute	Class	Membership		
			C1	C2	
x1	d1	C1	0.8	0.2	Discrete interval 1
x2	d1	C1	0.5	0.5	
x3	d2	C2	0.4	0.6	Discrete interval 2

(b) The first discretization scheme

Sample	Attribute	Class	Membership		
			C1	C2	
x1	d1	C1	0.8	0.2	Discrete interval 1
x2	d2	C1	0.5	0.5	Discrete interval 2
x3	d2	C2	0.4	0.6	

(c) The second discretization scheme

Fig. 3 A simple example illustrating that RFMD has advantages over rough set-based discretization methods



(a) Landsat 8 image (b) GF-2 image

Fig. 4 Area used for study

and (15), $Attribute\_C_1(x_1) = \inf\{0.8, 0.5\} = 0.5$, $Attribute\_C_1(x_2) = \inf\{0.8, 0.5\} = 0.5$, $Attribute\_C_1(x_3) = \inf\{0.4\} = 0.4$, $Attribute\_C_2(x_1) = \inf\{0.2, 0.5\} = 0.2$, $Attribute\_C_2(x_2) = \inf\{0.2, 0.5\} = 0.2$, $Attribute\_C_2(x_3) = \inf\{0.6\} = 0.6$, $Attribute^{-}C_1(x_1) = \sup\{0.8, 0.5\} = 0.8$, $Attribute^{-}C_1(x_2) = \sup\{0.8, 0.5\} = 0.8$, $Attribute^{-}C_1(x_3) = \sup\{0.4\} = 0.4$, $Attribute^{-}C_2(x_1) = \sup\{0.2, 0.5\} = 0.5$, $Attribute^{-}C_2(x_2) = \sup\{0.2, 0.5\} = 0.5$, $Attribute^{-}C_2(x_3) = \sup\{0.6\} = 0.6$, then $|Attribute\_C_1| = 0.5 + 0.5 + 0.4 = 1.4$, $|Attribute\_C_2| = 0.2 + 0.2 + 0.6 = 1.0$, $|Attribute^{-}C_1| = 0.8 + 0.8 + 0.4 = 2.0$, $|Attribute^{-}C_2| = 0.5 + 0.5 + 0.6 = 1.6$. According to (16), $\eta_1 = (1.4/2.0 + 1.0/1.6)/2 = 0.6625$. Similarly, in Fig. 3c, the equivalence classes under *Attribute* are $\{x_1\}$ and $\{x_2, x_3\}$, then $|Attribute\_C_1| = 0.8 + 0.4 + 0.4 = 1.6$, $|Attribute\_C_2| = 0.2 + 0.5 + 0.5 = 1.2$, $|Attribute^{-}C_1| = 0.8 + 0.5 + 0.5 = 1.8$, $|Attribute^{-}C_2| = 0.2 + 0.6 + 0.6 = 1.4$. According to (16), $\eta_2 = (1.6/1.8 + 1.2/1.4)/2 = 0.8730$. It can be seen that $\eta_2 > \eta_1$, that is, the accuracy of discretization scheme in Fig. 3c is higher than that in Fig. 3b, which is consistent with the conclusion drawn from the previous analysis.

(2) Rough set-based discretization methods use (3) as the evaluation standard of the system compatibility after discretization. In Fig. 3b, $|Attribute\_C_1| = |\{x_1, x_2\}| = 2$, $|Attribute\_C_2| = |\{x_3\}| = 1$, then $\gamma 1 = \gamma_{Attribute}(C_1) + \gamma_{Attribute}(C_2) = (2 + 1)/3 = 1$. Similarly, in Fig. 3c, $|Attribute\_C_1| = |\{x_1\}| = 1$, $|Attribute\_C_2| = |\varnothing| = 0$, then $\gamma 2 = \gamma_{Attribute}(C_1) + \gamma_{Attribute}(C_2) = (1 + 0)/3 = 0.3333$. Since $\gamma 1 > \gamma 2$, the discretization methods based on rough set will choose the discretization scheme in Fig. 3b, so the best discretization scheme cannot be obtained.

In summary, RFMD not only makes full use of the known knowledge base to generate rules as well as rough set-based discretization methods, but also fully considers the uncertainty caused by the fuzzy components in the data, so the samples with large internal component differences will not be classified into the same interval in the process of discretization, thereby obtaining a discretization scheme with higher precision.

Experiments

We introduce the experimental environment and datasets. Then, we compare the optimal breakpoint set obtained by RFMD algorithm with the discretization results of the current mainstream methods, mainly from the number of intervals, data consistency and classification accuracy.

Data source

The datasets used in this paper are as follows: (1) a Landsat 8 image from the northwestern region of

Table 1 Number of discrete intervals in each band of Landsat 8 image

Method	B1	B2	B3	B4	B5	B6	B7
RFMD	109	67	58	64	63	55	71
RS-GA	153	69	56	52	76	103	61
EDiRa	135	71	86	45	52	58	73
CVD	98	73	65	67	72	58	71
RLGA	120	67	65	52	63	55	71

Table 2 Number of data errors in Landsat 8 image

Method	Inconsistencies	Discrete intervals
RFMD	0	487
RS-GA	5	570
EDiRa	13	520
CVD	17	504
RLGA	2	493

Table 3 Number of discrete intervals in each band of GF-2 image

Method	B1	B2	B3	B4
RFMD	267	458	207	103
RS-GA	389	502	397	103
EDiRa	405	517	253	132
CVD	299	461	278	115
RLGA	267	461	247	103

Table 5 Number of discrete intervals in each attribute of the first methylation dataset

Method	Mean	Model prediction	Interpulse duration ratio
RFMD	210	189	138
RS-GA	244	269	156
EDiRa	241	296	34
CVD	205	229	129
RLGA	225	260	71

Zhejiang Province, China, and a GF-2 image from Lingshui County, Hainan Province, China, as shown in Fig. 4. Where, Landsat 8 satellite data contains seven bands, while GF-2 satellite data contains four bands [40]. In the experiment, the objects on Landsat 8 image were divided into seven categories: broadleaf, town, needles, farmland, lei bamboo, water, and moso bamboo; the objects on GF-2 image were divided into five categories: construction, bare land, farmland, vegetation, and water. (2) Two methylation datasets, including N6-methyladenine (6 mA) and N4-methylacytosine (4mC) [41, 42]. The three attributes of the first methylation dataset are: mean, model prediction, and interpulse duration ratio; the three attributes of the second methylation dataset are error, model prediction, and interpulse duration ratio. (3) The banknote authentication dataset extracted from banknote-like images [43] is divided into genuine banknote and forged banknote, and contains four attributes, namely variance, skewness, kurtosis, and entropy.

Configuration of experimental environment

In order to verify the effectiveness of the proposed method, all four algorithms were executed on a computer with Intel(R) Core (TM) i5-5200U CPU@2.20GHz processor, 12G RAM, and 512 g hard disk. Visualization, programming, simulation, testing and numerical calculation processing of this experiment were implemented in MATLAB (R2016a version) environment. Radiometric calibration of images, atmospheric correction, and comparison of results before and after discretization were performed under ENVI 5.3 environment.

Datasets

The ground reflection or emission spectral signal obtained by remote sensor is recorded by pixel. The interior of a pixel contains only one type, which is called pure pixel. However, in most cases, the interior of a pixel often contains many kinds of surface features, and this kind of pixel is called mixed pixel. The mixed pixel records the comprehensive spectral information of various types of ground objects. Several areas covering seven categories were randomly selected from Landsat 8 image and labeled. After integration, they were used as training samples to be discretized, with a total of 2621 samples. Among them, 308 cases were broadleaf, 245 were town, 322 were needles, 675 were farmland, 296 were lei bamboo, 262 were water, and 513 were moso bamboo. We used another group of samples with the same number of training samples as the test set. In the test set, 308 cases were broadleaf, 245 were town, 322 were needles, 675 were farmland, 296 were lei bamboo, 262 were water, and 513 were moso bamboo. Let N be the number of samples, and C be the number of categories, then the initial fuzzy segmentation matrix of the training set is:

$$PM^0 = \begin{bmatrix} f_1(x_1) & f_1(x_2) & \cdot & \cdot & \cdot & f_1(x_N) \\ f_2(x_1) & f_2(x_2) & \cdot & \cdot & \cdot & f_2(x_N) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ f_C(x_1) & f_C(x_2) & \cdot & \cdot & \cdot & f_C(x_N) \end{bmatrix} \tag{18}$$

Where,

$$f_j(x_i) = \begin{cases} 1, & x_i \text{ belongs to class } j \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

where $1 \le i \le N$ and $1 \le j \le C$

Table 4 Number of data errors in GF-2 image

Method	Inconsistencies	Discrete intervals
RFMD	0	1035
RS-GA	14	1391
EDiRa	25	1307
CVD	30	1153
RLGA	7	1078

Table 6 Number of data errors in the first methylation dataset

Method	Inconsistencies	Discrete intervals
RFMD	12	537
RS-GA	80	669
EDiRa	113	571
CVD	259	563
RLGA	71	556

Table 7 Number of discrete intervals in each attribute of the second methylation dataset

Method	Error	Model prediction	Interpulse duration ratio
RFMD	141	332	242
RS-GA	180	448	243
EDiRa	148	415	219
CVD	150	373	228
RLGA	143	363	216

In the beginning, the above segmentation matrix is substituted into (11) to initialize the cluster center of each category. Then, all the pixels contained in each band were sorted and de duplicated according to the brightness value, and the initial breakpoints of seven bands were 1314, 1517, 1056, 1211, 1086, 1920, and 1832, totaling 9936.

Similarly, in GF-2 image, there were 7554 training samples to be discretized. Among them, 2094 cases were construction, 775 cases were bare land, 1478 cases were farmland, 2251 cases were vegetation, and 956 cases were water. We used another group of samples with the same number of training samples as the test set. In the test set, 2094 cases were construction, 775 cases were bare land, 1478 cases were farmland, 2251 cases were vegetation, and 956 cases were water. All the pixels contained in each band were sorted and de duplicated according to the brightness value. The initial breakpoints of four bands were 3685, 3769, 2535, and 757, totaling 10,746. In the methylation datasets, there were 3709 training samples to be discretized. Among them, 1290 cases were 6 mA and 2419 cases were 4mC. A total of 1500 samples were tested. Among them, 500 cases were 6 mA and 1000 cases were 4mC. All the values contained in each attribute of the first methylation training set were sorted and de duplicated, and the initial breakpoints of three attributes were 1718, 1748, 960, totaling 4426. All the values contained in each attribute of the second methylation training set were sorted and de duplicated, and the initial breakpoints of three attributes were 564, 1748, 960, totaling 3272. In the banknote authentication dataset, there were 1072 training samples to be discretized. Among them, 562 cases were genuine banknotes and 510 cases were forged banknotes. A total of 300 samples were tested. Among them, there were 200 genuine banknotes and 100 forged banknotes. All the values contained in each attribute were sorted and de duplicated, and the initial breakpoints of four attributes were 1052, 996, 1015, 940, totaling 4003.

Our method was compared with RS-GA [24], EDiRa [16], CVD [17], and RLGA [18] mainly from the data consistency and the number of intervals. Finally, we trained the neural network classifier with the discretized samples of the above methods respectively, and verified the effectiveness of the proposed method by comparing the classification accuracy obtained by each method.

Table 8 Number of data errors in the second methylation dataset

Method	Inconsistencies	Discrete intervals
RFMD	0	715
RS-GA	6	871
EDiRa	11	782
CVD	15	751
RLGA	3	722

Table 9 Number of discrete intervals in each attribute of the banknote authentication dataset

Method	Variance	Skewness	Kurtosis	Entropy
RFMD	6	7	7	7
RS-GA	11	11	13	4
EDiRa	14	8	12	3
CVD	10	10	12	3
RLGA	6	8	8	8

Data consistency and number of breakpoints

The discretization results obtained on Landsat 8 image by RFMD, RS-GA, EDiRa, CVD, and RLGA are shown in Table 1 and Table 2.

It can be seen that the number of intervals obtained by RFMD algorithm is 487, which is the least in all algorithms, and there is no data error. The number of intervals in RS-GA algorithm is the largest in all algorithms, reaching 570, followed by EDiRa algorithm with 520. The number of data errors of these two algorithms are 5 and 13 respectively. The number of intervals of CVD algorithm is only 17 more than that of RFMD algorithm, but the number of data errors is the largest in all algorithms, with 17 errors. The number of intervals of RLGA is 493, and the number of data errors is 2, which is second only to RFMD. Table 3 and Table 4 show the results of the number of intervals in each band and data inconsistency obtained by RFMD, RS-GA, EDiRa, CVD, and RLGA on GF-2 image.

It can be seen that the number of intervals obtained by RFMD algorithm is 1035, which is the least in all

Table 10 Number of data errors in the banknote authentication dataset

Method	Inconsistencies	Discrete intervals
RFMD	0	27
RS-GA	1	39
EDiRa	2	37
CVD	3	35
RLGA	0	30

Table 11 Key differences among the mentioned discretization methods

Method	Direction	Attributes	Prior-knowledge	Uncertainty
RFMD	Evolutionary search	Multivariate	No need	Incompleteness & Fuzziness
RS-GA	Evolutionary search	Multivariate	No need	Incompleteness
EDiRa	Top-Down	Univariate	Need	Incompleteness
CVD	Bottom-Up	Univariate	Need	Incompleteness
RLGA	Evolutionary search	Multivariate	No need	Incompleteness

algorithms, and there is no data error. The number of intervals in RS-GA algorithm is the largest in all algorithms, reaching 1391, followed by EDiRa algorithm with 1307. The number of data errors of these two algorithms are 14 and 25 respectively. The number of intervals in CVD algorithm is 118 more than that in RFMD algorithm, and the number of data errors is the most in all algorithms, which is 30. RLGA has 1078 intervals and 7 data errors, which is second only to RFMD. Table 5 and Table 6 show the results of the number of intervals in each attribute and data inconsistency obtained by RFMD, RS-GA, EDiRa, CVD, and RLGA on the first methylation dataset.

It can be seen that the number of intervals obtained by RFMD algorithm is 537, which is the least in all algorithms, and the number of data errors is also the least in all algorithms, with 12. The number of intervals in RS-GA algorithm is the largest in all algorithms, reaching 669, followed by EDiRa algorithm with 571. The number of data errors of these two algorithms are 80 and 113 respectively. The number of intervals in CVD algorithm is 26 more than that in RFMD algorithm, and the number of data errors is the most in all algorithms, which is 259. RLGA has 556 intervals and 71 data errors, which is second only to RFMD. Table 7 and Table 8 show the results of the number of intervals in each attribute and data inconsistency obtained by RFMD, RS-GA, EDiRa, CVD, and RLGA on the second methylation dataset.

It can be seen that the number of intervals obtained by RFMD algorithm is 715, which is the least in all algorithms, and there is no data error. The number of intervals in RS-GA algorithm is the largest in all algorithms, reaching 871, followed by EDiRa algorithm with 782. The number of data errors of these two algorithms are 6 and 11 respectively. The number of intervals in CVD algorithm is 36 more than that in RFMD algorithm, and the number of data errors is the most in all algorithms, which is 15. RLGA has 722 intervals and 3 data errors, which is second only to RFMD. Table 9 and Table 10 show the results of the number of intervals in each attribute and data inconsistency obtained by RFMD, RS-GA, EDiRa, CVD, and RLGA on the banknote authentication dataset.

It can be seen that the number of intervals obtained by RFMD algorithm is 27, which is the least in all algorithms, and there is no data error. The number of intervals in RS-GA algorithm is the largest in all algorithms, reaching 39, followed by EDiRa algorithm with 37. The number of data errors of these two algorithms are 1 and 2 respectively. The number of intervals in CVD algorithm is 8 more than that in RFMD algorithm, and the number of data errors is the most in all algorithms, which is 3. RLGA has 30 intervals and no data error, which is second only to RFMD.

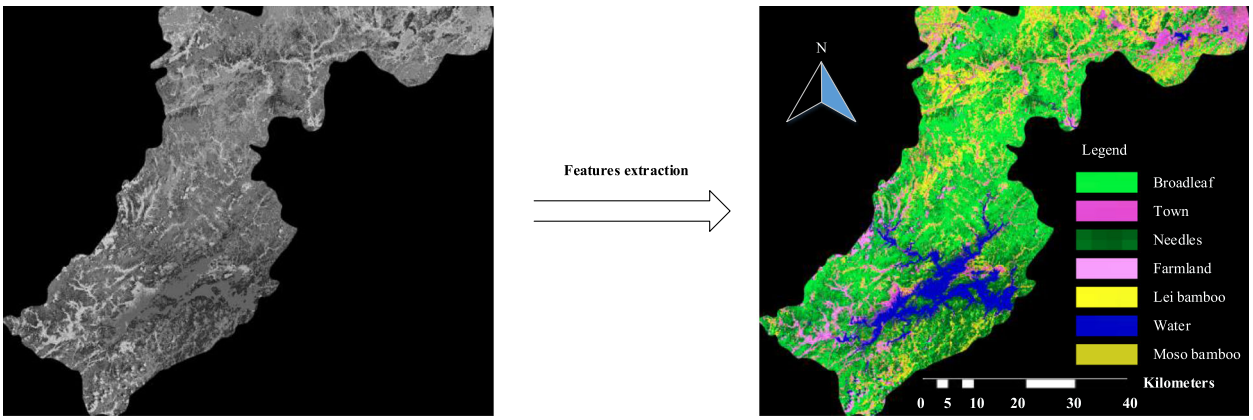
Although the discretization standards adopted by EDiRa and CVD have certain rationality, the relatively fixed partition criteria cannot comprehensively measure the discrete intervals. In addition, both EDiRa and CVD require the distribution information of sample attribute values in the dataset to improve the accuracy of interval partition. RS-GA adopts the discretization standard based on rough set, so it can achieve better results without any prior information. However, RS-GA lacks the ability to describe fuzzy components in the data, and its performance is often poor in the datasets with complex types. RLGA introduces reinforcement learning mechanism in crossover and mutation operations to improve the search efficiency of genetic algorithm. It keeps the data error at a low level and constantly seeks solutions with the least number of intervals. However, like RS-GA,

Table 12 Classification results in Landsat 8 image

Method	Overall accuracy	Kappa coefficient
RFMD	0.9428	0.9314
RS-GA	0.9275	0.9131
EDiRa	0.9222	0.9067
CVD	0.8993	0.8793
RLGA	0.9351	0.9223

Table 13 Classification results in GF-2 image

Method	Overall accuracy	Kappa coefficient
RFMD	0.9734	0.9655
RS-GA	0.9297	0.9083
EDiRa	0.9076	0.8795
CVD	0.8752	0.8385
RLGA	0.9314	0.9106



(a) Original image

(b) Classification effect map

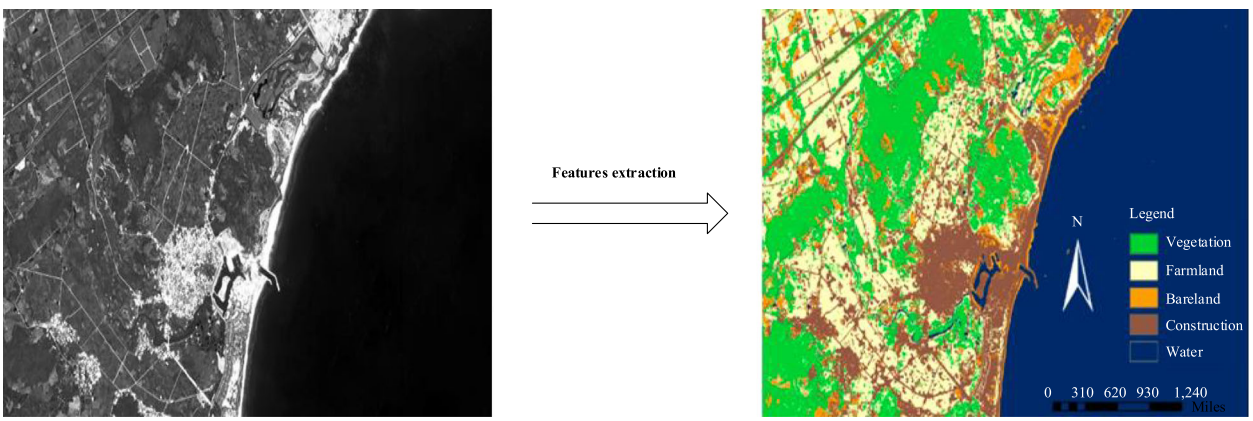
Fig. 5 Classification effect map of Landsat 8 image obtained by RFMD

the fitness function adopted by RLGA is only based on rough set and lacks the ability to describe the fuzzy components in the data. RFMD combines the advantages of rough set and fuzzy set, fully considers the fuzziness of data and the correlation among attributes, and determines the breakpoints in multiple continuous variables through evolutionary search. In this way, the performance of RFMD has been greatly improved and can adapt to most datasets with complex types. Therefore, the discretization result obtained by RFMD is the best in the five algorithms. The key differences among them are shown in Table 11.

Classification accuracy

We trained the neural network classifier with the discretized samples of these five algorithms, and obtained the classification results of Landsat 8 image and GF-2 image, as shown in Table 12 and Table 13.

It can be seen that the classification accuracy of RFMD is the best among the five algorithms. RS-GA, EDiRa, and RLGA have fewer data errors than CVD. Accordingly, RS-GA, EDiRa, and RLGA have higher classification accuracy than CVD. Figure 5 is a classification effect map of Landsat 8 image obtained by RFMD. It can be seen that the texture of the feature information in the figure is clear, the boundaries of different types of objects are more obvious, and there are almost no noise spots. The seven categories of broadleaf, town, needles, farmland, lei bamboo, water, and moso bamboo on the image can be effectively identified. Figure 6 is a classification effect map of GF-2 image obtained by RFMD. The texture of the feature information in the figure is clear, and the boundaries of different types of objects are very obvious. The five categories of construction, bare land, farmland, vegetation, and water on the image can be effectively identified.



(a) Original image

(b) Classification effect map

Fig. 6 Classification effect map of GF-2 image obtained by RFMD

Table 14 Classification results in the first methylation dataset

Method	Overall accuracy	Kappa coefficient
RFMD	0.9687	0.9308
RS-GA	0.9380	0.8626
EDiRa	0.9233	0.8310
CVD	0.9093	0.8031
RLGA	0.9453	0.8791

Tables 14, 15, and 16 show the classification results of the five algorithms on the first methylation dataset, the second methylation dataset, and the banknote authentication dataset, respectively. It can be seen that the classification accuracy of RFMD is the best in all algorithms. Therefore, the discretization scheme obtained by RFMD can achieve good results in classification accuracy.

Conclusion and future work

The data collected by edge nodes are often large in scale, complex in type, with incomplete, fuzzy, and other uncertain information. In order to lighten the system load, decrease the data inconsistency, and relieve the pressure on the centralized cloud, a discretization algorithm based on rough fuzzy model (RFMD) is proposed for intelligent data preprocessing of edge-cloud computing. The work of this paper mainly comes from the following aspects: (1) we create a fuzzy set for each category, and initialize all cluster centers according to the attribute values of the samples and the initial fuzzy segmentation matrix; (2) we use fuzzy *c*-means to obtain the membership function of each category, and establish the fitness function of genetic algorithm based on rough fuzzy model; (3) for each individual in the population, the average approximation accuracy of the corresponding discretization scheme is obtained by calculating the upper and lower approximations of all samples; (4) in each genetic operation, the fitness of all individuals in the population is calculated by the number of breakpoints and the average approximation accuracy of the discretization scheme, and the global variable is used to reserve the individual with the highest fitness, so as to obtain the optimal discretization scheme; (5) simulation experiments on real remote sensing datasets show that the proposed method can achieve good results in the number of discrete intervals, data consistency, and classification accuracy.

Table 15 Classification results in the second methylation dataset

Method	Overall accuracy	Kappa coefficient
RFMD	0.9633	0.9190
RS-GA	0.9247	0.8331
EDiRa	0.9100	0.8035
CVD	0.8960	0.7752
RLGA	0.9387	0.8643

Table 16 Classification results in the banknote authentication dataset

Method	Overall accuracy	Kappa coefficient
RFMD	0.9933	0.9851
RS-GA	0.9500	0.8872
EDiRa	0.9100	0.8010
CVD	0.8833	0.7494
RLGA	0.9767	0.9479

The future research work includes: (1) compare the performance of the proposed method on multiple classifiers to optimize the algorithm model, and further improve the efficiency of edge-cloud computing; (2) test and improve the proposed method in different datasets to expand its application scope, and further reduce the cost of data analysis and security management of edge cloud.

Acknowledgments

The authors would like to thank the support of the laboratory, university and government.

Authors' contributions

All authors take part in the discussion of the work described in this paper. The author(s) read and approved the final manuscript.

Authors' information

Qiong Chen received his B.Eng. degree at Beijing University of Posts and Telecommunications, P.R. China, 2007 and M.Eng. degree at Politecnico di Torino, Italy, 2012. Now he is a Ph.D. student at College of Information Science and Technology, Hainan University, P.R. China. His research interests include remote sensing image processing, evolutionary computing, granular computing, fuzzy decision-making, rough sets, big data analytics and multi-source data fusion.

Mengxing Huang received the Ph.D. degree from Northwestern Polytechnical University,Xi'an, China, in 2007. He then joined staff with the Research Institute of Information Technology, Tsinghua University as a Postdoctoral Researcher. In 2009, he joined Hainan University. He is currently a Professor and a Ph.D. Supervisor of computer science and technology, and the Dean of School of Information and Communication Engineering. He is also the Executive Vice-Presedent of Hainan Province Institute of Smart City, and the Leader of the Service Science and Technology Team with Hainan University. He has authored or coauthored more than 60 academic papers as the first or corresponding author. He has reported 12 patents of invention, owns 3 software copyright, and published has 2 monographs and 2 translations. He has been awarded Second Class and Third Class Prizes of The Hainan Provincial Scientific and Technological Progress. His current research interests include signal processing for sensor system, big data, and intelligent information processing.

Funding

This work was supported by Hainan Provincial Natural Science Foundation of China (Grant #: 2019CXTD400), the National Key Research and Development Program of China (Grant #: 2018YFB1404400). (Corresponding author: Mengxing Huang.)

Availability of data and materials

The Landsat 8 and GF-2 datasets, the methylation datasets, and the banknote authentication dataset used to support the findings of this study are included within the article. All the data and materials in this article are real and available.

Competing interests
The authors declare no conflict of interest. The sponsors had no role in the design, execution, interpretation, or writing of the study. These no potential competing interests in our paper. And all authors have seen the manuscript and approved to submit to your journal. We confirm that the content of the manuscript has not been published or submitted for publication elsewhere.

Author details
[1]State Key Laboratory of Marine Resource Utilization in South China Sea, Haikou 570228, China. [2]School of Information and Communication Engineering, Hainan University, Haikou 570228, China.

Received: 1 September 2020 Accepted: 26 November 2020
Published online: 18 January 2021

References

1. Taleb T, Samdanis K, Mada B et al (2017) On multi-access edge computing: a survey of the emerging 5G network edge cloud architecture and orchestration. IEEE Commun Surveys Tutorials 19(3):1657–1681
2. Pan J, Mcelhannon J (2018) Future edge cloud and edge computing for internet of things applications. IEEE Internet Things J 5(1):439–449
3. Fernando N, Loke SW, Rahayu W et al (2019) Computing with nearby Mobile devices: a work sharing algorithm for Mobile edge-clouds. IEEE Transact Cloud Comput 7(2):329–343
4. Rodrigues TG, Suto K, Nishiyama H et al (2017) Hybrid method for minimizing service delay in edge cloud computing through VM migration and transmission power control. IEEE Trans Comput 66(5):810–819
5. Wu H, Li X, Deng Y (2020) Deep learning-driven wireless communication for edge-cloud computing: opportunities and challenges. J Cloud Comp 9:21 (2020)
6. Jarray A, Karmouch A, Salazar J et al (2017) Efficient resource allocation and dimensioning of media edge clouds infrastructure. J Cloud Comp 6:27 (2017)
7. Liu H, Eldarrat F, Alqahtani H et al (2018) Mobile edge cloud system: architectures, challenges, and approaches. IEEE Syst J 12(3):2495–2508
8. Garcia S, Luengo J, Saez JA et al (2013) A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Trans Knowl Data Eng 25(4):734–750
9. Chen Q, Huang M, Wang H et al (2018) A Feature Preprocessing Framework of Remote Sensing Image for Marine Targets Recognition. In: 2018 OCEANS - MTS/IEEE Kobe techno-Oceans (OTO), pp 1–5
10. Simon HA (1996) The sciences of the artificial, 3rd edn. MIT Press, Cambridge
11. Dbouk T, Mourad A, Otrok H et al (2019) A novel ad-hoc Mobile edge cloud offering security services through intelligent resource-aware offloading. IEEE Trans Netw Serv Manag 16(4):1665–1680
12. Liu J, Wu J, Sun L et al (2020) Image data model optimization method based on cloud computing. J Cloud Comp 9(1):1
13. Ramirezgallego S, Garcia S, Mourinotalin H et al (2016) Data discretization: taxonomy and big data challenge. Wiley Interdisciplin Rev Data Mining Knowl Discov 6(1):5–21
14. Chlebus BS, Nguyen SH (1998) On finding optimal Discretizations for two attributes. Lect Notes Comput Sci:537–544
15. Wong AK, Chiu D (1987) Synthesizing statistical knowledge from incomplete mixed-mode data. IEEE Trans Pattern Anal Mach Intell 9(6):796–805
16. De Sa CR, Soares C, Knobbe A et al (2016) Entropy-based discretization methods for ranking data. Inform Sci 329:921–936
17. Wu B, Zhang L, Zhao Y et al (2014) Feature selection via Cramer's V-test discretization for remote-sensing image classification. IEEE Trans Geosci Remote Sens 52(5):2593–2606
18. Chen Q, Huang M, Xu Q et al (2020) Reinforcement learning-based genetic algorithm in optimizing multidimensional data discretization scheme. Math Probl Eng 2020(1):1–13
19. Nguyen SH, Skowron A (1995) Quantization of real value attributes-rough set and Boolean reasoning approach. In: Proc. second joint Ann. Conf. Information sciences (JCIS), pp 34–37
20. Kara N, Soualhia M, Belqasmi F et al (2014) Genetic-based algorithms for resource management in virtualized IVR applications. J Cloud Comp 3:15
21. Nikravesh AY, Ajila SA, Lung C (2018) Using genetic algorithms to find optimal solution in a search space for a cloud predictive cost-driven decision maker. J Cloud Comp 7:20
22. Chen C, Li Z, Qiao S et al (2003) Study on discretization in rough set based on genetic algorithm. In: International conference on machine learning and cybernetics, pp 1430–1434
23. Ren ZH, Hao Y, Wen B et al (2011) A heuristic genetic algorithm for continuous attribute discretization in rough set theory. Adv Mater Res 2011: 132–136
24. Dai J (2004) A genetic algorithm for discretization of decision systems. In: International conference on machine learning and cybernetics, pp 1319–1323
25. Ishibuchi H, Yamamoto T, Nakashima T (2001) Fuzzy data mining: effect of fuzzy discretization. In: Proc. IEEE Int'l Conf. Data Mining (ICDM), pp 241–248
26. Krinidis S, Chatzis V (2010) A robust fuzzy local information C-means clustering algorithm. IEEE Trans Image Process 19(5):1328–1337
27. Saltos R, Weber R, Maldonado S et al (2017) Dynamic rough-fuzzy support vector clustering. IEEE Trans Fuzzy Syst 25(6):1508–1521
28. Dougherty J, Kohavi R, Sahami M et al (1995) Supervised and unsupervised discretization of continuous features. In: International conference on machine learning. Elsevier, pp 194–202.
29. Goldberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley Professional, USA
30. Ramirezgallego S, Garcia S, Benitez JM et al (2016) Multivariate discretization based on evolutionary cut points selection for classification. IEEE Trans Cybern 46(3):595–608
31. Pawlak Z (1992) Rough sets: theoretical aspects of reasoning about data. Kluwer Academic Publishers, Norwell
32. Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353
33. Mitra S, Banka H, Pedrycz W (2006) Rough–fuzzy collaborative clustering. IEEE Trans Syst Man Cybern B Cybern 36(4):795–805
34. Han Y, Shi P, Chen S (2015) Bipolar-valued rough fuzzy set and its applications to the decision information system. IEEE Trans Fuzzy Syst 23(6): 2358–2370
35. Dash S, Luhach AK, Chilamkurti N et al (2019) A Neuro-fuzzy approach for user behaviour classification and prediction. J Cloud Comp 8:17 (2019)
36. Ismaeel S, Karim R, Miri A (2018) Proactive dynamic virtual-machine consolidation for energy conservation in cloud data centres. J Cloud Comp 7:10 (2018)
37. Elrawy M, Awad A, Hamed H (2018) Intrusion detection systems for IoT-based smart environments: a survey. J Cloud Comp 7:21
38. Jin R, Yuri B, Chibuike M (2009) Data discretization unification. Knowl Inf Syst 19(1):1–29
39. Huang M, Chen Q, Wang H (2020) A multivariable optical remote sensing image feature discretization method applied to marine vessel targets recognition. Multimed Tools Appl 2020:4597–4618
40. Wu D, Huang M, Zhang Y, Bhatti UA, Chen Q (2018) Strategy for assessment of disaster risk using typhoon hazards modeling based on chlorophyll-a content of seawater. EURASIP J Wirel Commun Netw 2018(1)
41. Xiao C, Zhu S, He M et al (2018) N6-Methyladenine DNA modification in the human genome. Molecularcell 71(2):306–318
42. Yuan D, Xing J, Luan M et al (2020) DNA N6-methyladenine modification in wild and cultivated soybeans reveal different patterns in nucleus and cytoplasm. Front Genet. https://doi.org/10.3389/fgene.2020.00736
43. Li Y, Huang M, Zhang Y et al (2020) Automated Gleason grading and Gleason pattern region segmentation based on deep learning for pathological images of prostate cancer. IEEE Access 8:117714–117725

Publisher's Note