TELKOMNIKA, Vol. 11, No. 10, October 2013, pp. 5782 ~ 5788
ISSN: 2302-4046

A Clustering Algorithm Based on Rough Set and Genetic Algorithm

Yushu Xiong
Department of Electronic Engineering and Automation, Chongqing Vocational Institute of Engineering
400037 Chongqing, China
e-mail:ysxiongyeah@163.com

Abstract

With the development of computer and information technology, the capacity of data and information is increasing. The processing of data and information becomes the hot issue in the current scientific community. Rough set and genetic algorithm are two data mining and processing technologies which had been commonly used. Rough set can process data quickly and the algorithm is simple. The convergence of genetic algorithm is fast and the robustness is good. This paper puts forward the effective clustering algorithm based on the combined control of rough set and genetic algorithm, then does simulation experiment of segmentation images by numerical simulation based on matlab programming, finally gets the curve of calculation process and the effect diagram of image segmentation , verifies the effectiveness of the algorithm.

Keywords: *Rrough Set, Genetic Algorithm, Clustering Algorithm, Data Mining, Image Processing;*

1. Introduction

With the increasing amount of data information, how to extract useful information from complex data and handle the data rapidly and efficiently is the focus of data processing algorithms. According to this, this article introduces effective clustering algorithm controlled by rough set genetic algorithm [1]. The FRC control strategy algorithm of the rough set makes algorithm simpler and easier to complete and fast and effective. GA test functions using genetic algorithms makes the algorithm has good convergence and robustness. Typically, people have obtained information from around the world that is often inaccurate, incomplete or uncertain. However, people still have to rely on this information, through thinking and summary, to make a decision. This means that people should be able to deal with this uncertain information before drawing a correct conclusion and making a reasonable decision.

These uncertain information need to rely on fuzzy sets and rough sets theory as analysis tool, especially all kinds of complicated data. Rough set and fuzzy set make up non precise information to express two important drawbacks, which are respectively indiscernible and ambiguous. The former is the essential attribute of things; the latter is a classification problem. Fuzzy set theory is proposed by Zadeh in 1965, it has been proved practical in chemistry and other disciplines. In contrast, Pawlak introduced the rough set theory in 1985, although its theory is very popular in many disciplines, it isn't mentioned in the chemical. This also shows that both the rough set theory and the other set theory have essential difference [2].

The traditional set theory, such as fuzzy sets, the elements in the collection can be clearly expressed. Using membership function describes common elements and collection property relations, property relations can be with or without. The definition of the membership function does not take into account the elements' uncertain problems in the collection, in order to deal with uncertainty, fuzzy sets are proposed. The fuzzy set membership function, its value can be from the closed interval of 0 to 1, and allowing the segment [3,4]. The fuzzy set membership function describes the events on the extent to occurs, rather than whether it occurs.

However, the element in the rough set theory isn't mainly concept. The rough set represents different mathematical methods of treatment ambiguity and uncertainty. The rough set theory contains people awareness of things and perception on the definition of the set. In

Received April 24, 2013; Revised July 2, 2013; Accepted July 16, 2013

other words, people have seen useful information elements part, ignoring the same place of the two elements, and find similarities. The rough set theory is particularly suitable for inaccurate or incomplete data reasoning, to discover the hidden patterns and rules; its application field continues to expand.This paper establishes mathematical model of rough set and genetic algorithm clustering algorithm and introduces the mathematical expressions of similar matrix and sample clustering center and density. At last, this paper establishes the validation procedure of clustering algorithm through the form of MATLAB program [5, 6]. Then, it divides an image and gets the segmentation effect diagram which proves the validity and rationality of the algorithm.

2. Effective Clustering Algorithm Controlled by Rough Set and Genetic Algorithm

In the massive process of data mining, some of the data are often vague and cannot be classified. These data cannot be counted in a collection and cannot be existed in subset and the complement of a subset, but it can be counted in the boundary of the set. The study of rough set can use the similarity matrix which can be expressed as follows:

$$\begin{matrix} 1 & a_{12} & \cdots & a_{1i} \\ a_{21} & 1 & \cdots & a_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ a_{j1} & a_{j2} & \cdots & 1 \end{matrix} \quad (1)$$

In formula (1), a_{ji} is the value of object j according to a certain degree of similarity of the object i. If the value of a_{ji} is larger, the similarity between two objects is smaller.

The use of rough sets can be automatically controlled to extract data. This paper presents a new control strategy FRC--fuzzy-rough control. The basic idea of this control strategy is: We can use a data recording manner to record the state and representative measures in the control process and integrate these data using rough sets. We can summarize the data integration process as follows [7-9]:

The rule one IF Condition 1 corresponds to THEN TAKE project 1;

The rule second IF Condition 2 corresponds to THEN TAKE project 2;

The rule third IF Condition 3 corresponds to THEN TAKE project 3.

This data processing strategy is called paradigm learning. This approach based on the coarse control and fuzzy control. Rough control has the features of simple, quick and easy. Another feature is data control algorithm comes from the data itself. Its decision-making and reasoning process is easier. It is easier than fuzzy control to inspect and operate and it is applied in a simple algorithm [10].

The genetic algorithm is an information data processing method based on the principle of binary data and genetic information. GA genetic test functions can be added in the clustering algorithm for the convergence of clustering algorithm. The GA test function is shown as follows [11]:

$$p(y_j) = 0.02 + \sum_{i=1}^{25} \frac{1}{i + \sum_{j=1}^{2} (y_j - a_{ji})^6} \quad (2)$$

In formula (2), a is judgment matrix. Function has more than one maximum. Generally speaking, if the function value is greater than 1 it is convergence. This test method is fast and robust performance is good.

Assuming that data samples is $b_k = (b_{k1}, b_{k2}, ...b_{kn})$, the distance between b_k and b_l can be defined as:

$$d(b_k, b_l) = \sqrt{\sum_{z=1}^{c}(b_{kz} - b_{lz})}, k, l = 1,2,...n \tag{3}$$

The density of the data samples at the point of b_k can be expressed as [12]:

$$M_k = \sum_{l=1}^{n} \frac{d(b_k, b_l)}{\sum_{k=1}^{n} d(b_l, b_e)}, k = 1,2,...n \tag{4}$$

Then, we can select the next best cluster center. First, we can make the sample density distribution around more intensive. The density can be defined is:

$$Mp_k = \exp(-\frac{M_k}{\sum_{i=1}^{g} d(B_j, s_i)}), j = 1,2,...n, j \neq i \tag{5}$$

The calculation flow chart based on rough sets and genetic algorithm is shown in Figure 1.

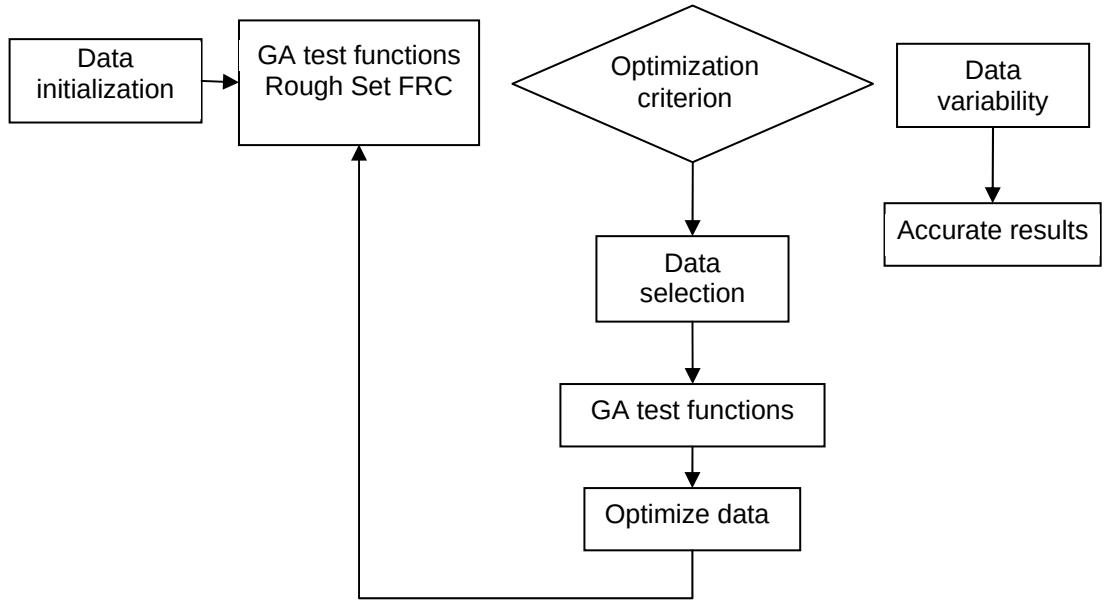

Figure 1. The effective clustering algorithm flowchart

As is shown in Figure 1, data set after the initialization can enter into optimized Data Processing through the control of the GA genetic algorithm test functions and rough set FRC. The data after optimization can meet the optimization criteria and can export the accurate results. If it does not meet the optimization criteria, we need to reintegrate and calculate the data through data optimization of genetic algorithm.

3. Application Examples on Rough Set Rule Reduction Algorithm

Rough set theory is applied to the composition of quantitative analysis problem, to model data source that is a protein component, wherein the amino acid of 10 coded has 5 attributes: a1=PIE is expressed as side-chain fat-soluble; a2=DGR=ΔG is said that the protein convert to water content; a3=SAC is surface area; a4=MR is molecular refraction index; a5=LAM is side-chain polarity.

First of all, to establish decision table, then the quantitative attributes of {a1, a2, a3, a4, a5} and table decision attribute {d} constitute a decision table, which is shown in Table 1.

Table 1. A protein component of decision table

U	a1	a2	a3	a4	a5	d
X0	0.23	0.54	251.2	2.215	0.02	8.3
X1	0.48	0.51	296.5	2.966	0.03	8.2
X2	0.61	1.20	287.9	2.994	1.08	8.6
X3	0.45	1.40	282.9	2.933	0.11	10.3
X4	0.11	0.29	335.0	3.458	1.19	6.5
X5	0.51	0.76	311.6	3.243	1.43	8.8
X6	0.00	0.19	315.6	2.932	1.03	7.1
X7	0.15	0.25	337.2	3.856	1.06	7.9
X8	1.20	2.10	322.6	3.350	0.04	9.9
X9	1.28	2.00	324.0	3.518	0.12	8.7

The condition attributes are encoded as 3 quantitative intervals, such as using 1, 2 and 3 respectively expresses low, medium and high, all the attributes use natural number coding to quantitative interval, as shown in table 2.

Table 2. The properties of quantitative analysis

Property	Coding			
	1	2	3	4
a1	<0.115	[0.115,0.54)	[0.54,1.195)	>1.195
a2	<1.167	[1.167,0.75)	[0.75,1.25)	>1.25
a3	<264.5	[264.5,321.5)	[321.5,361.2)	>361.2
a4	<2.75	[2.75,3.54)	[3.54,4.861)	>4.861
a5	<1.21	[1.21,1.64)	[1.64,2.15)	>2.15
d	<8.31	[8.31,12.66)	[12.66,13.98)	

All the attribute set of each molecule is upper and lower approximation, to calculate all approximate accuracy that approximately equal to 100%. Therefore, the quality of classification is also approximately 100%.

The next step of rough set analysis is to build the minimal subsets of independent attributes, to ensure the quality of classification and collection has the same effect. There are four D reductions that are given as follows:

Set #1= {a2, a4}
Set #2= {a2, a7}
Set #3= {a2, a5, a1}
Set #4= {a2, a5, a6}

The intersection of all D reduction is the core of the property D. In this example, the D core is {a2}, which means that the property is the most important basis for classification. In order to ensure the classification quality is not reduced, this property can't be reduction. The base of reduction can be 2 or 3, the excess is not necessary, to remove the classification quality will not be affected, minimal reduction set is #1and #2. Therefore, the initial coding of information system attribute can be reduced from 7 to 2. This shows that based on the D core and D reduction, the relevant attributes can be further reduced. After the reduction, the information

system can be viewed as a decision table. The classification accuracy of reduction set #1 is shown in Table 3.

Table 3. The classification accuracy set {a2, a4}

Category	Accuracy		
	{a2, a4}	{a2}	{a4 }
1	1.000	0.361	0.162
2	1.000	0	0
3	1.000	0	0

It can be seen from the Table4 , the minimum reduction set #1={a2, a4} can ensure the classification accuracy that is 100%, while using separately a2 or a4 can't guarantee the classification accuracy, it shows that this algorithm is effective and correct.

4. Matlab Numerical Simulation Experiment of Efficient Clustering Algorithm

MATLAB mathematical software is professional programming experiment software. According to the mathematical model and flowchart of the first part, this paper establishes MATLAB program. The main programming steps are shown as follows [13-15]:

Step1: We can do the binary encoding of the clustering data;

Step2: We can assume that the number of iterative evolution is t and generate p data clustering groups;

Step3 : We can convert the binary data to decimal numbers and calculate clustering center based on the features of rough set;

Step4: We can input GA test function and calculate the fitness function;

Step5: We can filter clustering groups according to adaptation function and get the new data clustering group;

Step6: We can determine whether it comply with the principle of data optimization;

Step7: If it meets the principles, we can output the data. If not, we can enter the t+1-th time to select;

Step8: We can output clustering data groups;

This paper selects a landscape painting and divides the color of the image through the model of pixel segmentation. Figure 2 is the curves of MATLAB in the calculation process.

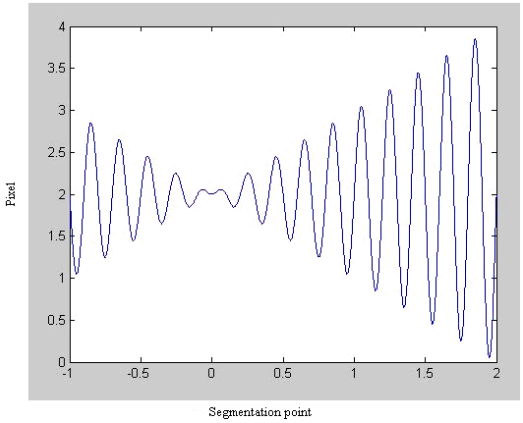

Figure 2. MATLAB calculation process curves

This paper selects several sets of image data groups to process the image segmentation and according to MATLAB gets the calculation time of different data groups which is shown in Table 4.

Table 4. Time comparison table of algorithms

Data set	Clustering algorithm time	Ordinary algorithm time
1	0.082	0.101
2	0.123	0.152
3	0.145	0.162
4	0.167	0.187

Table 5. Algorithms accuracy comparison table

Data set	Clustering algorithm time	Ordinary algorithm time
1	86.4	82.3
2	76.1	71.1
3	78.9	77.2
4	65.2	55.6

From the Table 4, we can see that, the speed of clustering algorithm in this article is higher than the ordinary algorithm. The highest speed of it is 0.082 seconds.
In table 5, we can see that, the accuracy of clustering algorithm in this article is higher than ordinary algorithm [16]. The highest accuracy is 86.4%. This paper selects a color image to process the image color clustering segmentation. Figure 3 is the selected color Landscape.

Figure 3. Original image segmentation

Figure 4. Image gray value segmentation results

Figure 5. Image segmentation results

Figure 4 is image gray value segmentation map through effective clustering algorithm gray pixels split. From it, we can see that the outline of gray pixels is obvious which proves the effectiveness of the clustering algorithm [17].

According to rough sets and genetic algorithms of this article, the final results of MATLAB color image segmentation are shown in Figure 5. It divided the image into four categories: black, white, dark gray and light gray. The boundary is very clear which proves the effectiveness of the algorithm.

5. Conclusion

This paper proposed a new effective clustering algorithm according to two advanced data processing methods--rough sets and genetic algorithms. The first part introduced rough sets and genetic algorithms in detail and introduced similarity matrix of rough set. It established clustering control mathematical model of rough set FRC according to freedom extract principle of the roughness data. It inserted GA test function into the clustering algorithm according to the characteristics of the genetic algorithm which improves the speed and convergence of data processing and increases the robustness of the clustering algorithm. The second part tested the effectiveness of the clustering algorithm through the professional data processing software MATLAB. It selected a landscape color image to process color clustering segmentation and got the segmentation effect diagram. From the calculation data, we can got that clustering algorithm of this article can improve the speed and accuracy of the data processing. The fastest speed is 0.082 seconds and the maximum accuracy is 86.4%. It also clustered the image as four categories: black, white, gray and light gray and got the obvious outline which proves the reliability of the algorithm.

References

[1] Kuang Yu Huang. An enhanced classification method comprising a genetic algorithm, rough set theory and a modified PBMF-index function. *Applied Soft Computing*. 2012; 12(1): 46-63.

[2] Xiaoyi Deng. An Enhanced Artificial Bee Colony Approach for Customer Segmentation in Mobile E-commerce Environment. *International Journal of Advancements in Computing Technology*. 2013; 5(1): 139-148.

[3] Youshyang Chen. Classifying credit ratings for Asian banks using integrating feature selection and the CPDA-based rough sets approach. *Knowledge-Based Systems*. 2012; 26(1): 259-270.

[4] Rafael Bello, José Luis Verdegay. Rough sets in the Soft Computing environment. *Information Sciences*. 2012; 212(1): 1-14.

[5] Georg Peters, Fernando Crespo, Pawan Lingras. Soft clustering – Fuzzy and rough approaches and their extensions and derivatives. *International Journal of Approximate Reasoning*. 2012; 54(2): 307-322.

[6] Hong Zhang, Suqun Cao. Fuzzy C-Means Clustering Algorithm Based Analysis Method for the Structure of Teachers in Local Engineering Colleges. *International Journal of Digital Content Technology and its Applications*. 2013; 7(3): 215-220.

[7] Hia Jong Teoh, Ching-Hsue Cheng, Hsing-Hui Chu. Fuzzy time series model based on probabilistic approach and rough set rule induction for empirical research in stock markets. *Data & Knowledge Engineering*. 2010; 567(1): 103-117.

[8] WC Chen, Ni-Bin Chang, Jeng-Chung Chen. Rough set-based hybrid fuzzy-neural controller design for industrial wastewater treatment. *Water Research*. 2010; 37(1): 95-107.

[9] Chihhua Hsu. Alternative rule induction methods based on incremental object using rough set theory. *Applied Soft Computing*. 2013; 13(1): 372-389.

[10] Huoyang Lin. Study on The Clustering Algorithm Based on The Sector-Ring Staggered Wireless Sensor Networks. *International Journal of Digital Content Technology and its Applications*. 2013; 7(3): 241-247.

[11] Jianhua Dai, Qing Xu. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Applied Soft Computing*. 2013; 13(1): 211-221.

[12] Mohammad Lutfi Othman, Ishak Aris, Mohammad Ridzal Othman. Rough-Set-and-Genetic-Algorithm based data mining and Rule Quality Measure to hypothesize distance protective relay operation characteristics from relay event report. *International Journal of Electrical Power & Energy Systems*. 2011; 33(8): 1437-1456.

[13] Yanyan Wang, Yanning Wang, Di Wu, Jiadong Ren. An Incremental Rapid DBSCAN Clustering Algorithm for Detecting Software Vulnerabilities. *Journal of Convergence Information Technology*. 2013; 8(3): 627-633.

[14] Pradipta Maji, Sushmita Paul. Rough set based maximum relevance-maximum significance criterion and Gene selection from microarray dat. *International Journal of Approximate Reasoning*. 2011; 52(3): 408-426.

[15] Huiling Chen, Bo Yang, Jie Liu, Dayou Liu. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. *Expert Systems with Applications*. 2011; 38(7): 9014-9022.

[16] Meng Fanrong, Gao Chunxiao, Liu Bing. Fuzzy Possibilistic Support Vector Machines for Class Imbalance Learning. Journal of Convergence Information Technology. 2013; 8(3): 692-701.

[17] Wu Deng, Wen Li, Xinhua Yang. A novel hybrid optimization algorithm of computational intelligence techniques for highway passenger volume prediction. *Expert Systems with Applications*. 2011; 38(4): 4198-4205.