Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 1498 – 1503

Procedia Engineering

www.elsevier.com/locate/procedia

Efficient Clustering Method Based on Rough Set and Genetic Algorithm

Jianyong Chen Changsheng Zhang*

College of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou, 325035, China

Abstract

In the process of traditional hard clustering, the obtained data objects in clusters are certain. However, the objects in different classes do not have clear boundaries between in reality. A method of dealing with uncertain boundary objects is provided by Rough set theory. Therefore, combing two methods of rough set theory and k-means cluster the objects. At the same time, though the traditional k-means algorithm has powerful local search capability, it easily falls into local optimum. The genetic algorithm can get the global optimal solution, but its convergence is fast. So in the process of clustering, rough set theory and genetic algorithm are introduced. An efficient clustering method based on rough set theory and genetic algorithm is provided. Finally, the experimental results show that the proposed algorithm has the ability to adjust the results and obtain the higher accuracy rate.

Keywords: Clustering; K-means; Rough set; Genetic Algorithm

1. Introduction

Clustering is one of the important researches in data mining[1,2]. Clustering is the process of dividing the set of data objects into several clusters consisting of similar objects. The obtained clusters from clustering are a set of data objects, these objects are similar to the objects in one cluster, but they are different from the objects in different clusters. The methods of clustering are hard clustering and vague clustering[3,4]. The traditional k-means algorithm belongs to hard clustering, in this way, a data object is divided into a certain cluster strictly, and it has no relationship with other clusters. The existing k-means algorithm is simple, its convergence is fast and its local search capability is powerful. But it has the disadvantage of falling into local optimum easily. Genetic algorithm is a global optimum algorithm based on the principle of biological evolution. It has two conspicuous features: implicit parallelism and global optimization. The fist one reflects the larger searching zone by only inspecting a small quantity structure,

* Corresponding author
E-mail address: jsj_zcs@126.com.

1877-7058 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.proeng.2011.08.278

which is easy to deal with; the latter one make the genetic algorithm have stronger robustness, which can avoid local optimum.

Therefore, in this paper, a new improved strategy is provided. That is to combine the k-means algorithm which has the strong local search ability with the genetic algorithm which has the strong global optimum. Making use of rough set theory is to deal with imprecise and incomplete knowledge[5,6]. Furthermore, the set of upper approximation, lower approximation and boundary to deal with the border objects. Through appropriate regulation and by playing their respective advantages, the clustering accuracy is improved. The experimental results show that the improved clustering algorithm is better for clustering.

2. Related Work

2.1. Rough set

Definition1. Given a decision table $S=(U,C,D,V,f)$, where $U=\{x_1,x_2,...,x_n\}$ is the non-empty limited set of objects, also called domain; where the set of attributes $A=C\cup D$, and $C\cap D=\varnothing$. Where C is a set of condition attribute, D is a set of decision attribute. $V=\underset{a\in C\cup D}{\cup}V_a$, that is, where V_a is the value range of attribute a, $f:U\times C\cup D\to V$ is an information function, that is $\forall a\in C\cup D$, $x\in U$, $f(x,a)\in V_a$ holds. With each attribute subset $B\subseteq(C\cup D)$, there is a binary indiscernibility relation $IND(B)=\{(x,y)\in U\times U\mid \forall a\in B, f(x,a)=f(y,a)\}$, $IND(B)$ is in short U/B. In the decision table, if $\forall a\in C$, $f(x,a)=f(y,a)$, there exists $f(x,D)\neq f(y,D)$, then the two objects are consistent, or the two objects are inconsistent.

Definition2. For a decision table $S=(U,C,D,V,f)$, $\forall B\subseteq C, X\subseteq U$, denote $U/B=\{B_1, B_2,\cdots,B_t\}$, then $B_-(X)=\cup\{B_i\mid B_i\in U/B, B_i\subseteq X\}$ is called lower approximation of X with respect to B, then $B^-(X)=\cup\{B_i\mid B_i\in U/B, B_i\cap X\neq\phi\}$ is called upper approximation of X with respect to B.

Definition3. For a decision table $S=(U,C,D,V,f)$, let $U/D=\{D_1,D_2,\cdots,D_n\}$ be the partition of D with respect to U, $U/P=\{P_1,P_2,\cdots,P_m\}$ be the partition of $P(P\subseteq C)$ with respect to U, $POS_P(D)=\underset{D_i\in U/D}{\cup}P_-(D_i)$ is called positive region P with respect to D, $U-POS_P(D)$ is called negative region P with respect to D.

Definition4. For a decision table $S=(U,C,D,V,f)$, $\forall b\in B\subseteq C$, if $POS_B(D)=POS_{B-\{b\}}(D)$, then b is not necessary for B with respect to D; Otherwise, then b is necessary for B with respect to D. For $\forall B\subseteq C$, if every element in B with respect to D is necessary, then B with respect to D is independent.

2.2. Genetic Algorithm

The heuristic algorithm is local search and its convergence is fast, but it is easy to fall into local optimum. The final result is only sub-optimal solution. Different methods are used to optimize the sub-optimal solution in paper [4-7]. In this paper, the fitness function of genetic algorithm is optimized

globally by rough set theory. The basic genetic algorithm in algorithms has much improvement, the following algorithms detail as follows:

1). The coding of chromosome

In this paper, the binary encoding is introduced, at the same time, the concept of approximation and lower approximation are introduced to code rough clustering. And the coding strategy is as follows: If the objects in the data set belong to the boundary or negative area in clusters, then the corresponding code of chromosome string is 1, and otherwise it is 0. Since binary encoding has the feature of simple, cross-compile and convenient, in the whole, genetic algorithm in easy to operate.

2). The initial population

The initial population is an important aspect of genetic algorithms. The characteristics of the initial population on the results and efficiency have an important effect. To achieve the global optimum, the initial population in the solution space should be distributed. Standard genetic algorithm is randomly generated the initial population, which may lead to the initial population distribute uneven in the solution space, thus affecting the performance of the algorithm. 50% of initial populations are from the generated seeds of the heuristic algorithm, 50% of the initial populations are from the randomly generated seed. This has reached the purpose of seed diversity.

3). The fitness function

In the genetic algorithm, the fine level of the optimal solution can be achieved by making use of the fitness function which is to measure each individual's optimization calculation in the groups. In this paper, the fitness function is constructed as follows: $F(v)=\beta\cdot f(x)+p(x)=\beta\times(1-\frac{card(x)}{card(C)})+\frac{card(POS_X(D))}{card(POS_C(D))}$. Where $f(x)=(1-\frac{card(x)}{card(C)})$ indicates the chromosome x that are not included in the proportion of condition attributes. $\beta=\frac{1}{card(POS_C(D))}$ ensures a superset of computing attribute reduction is superior to the decrease of the number of condition attributes. $p(x)=\frac{card(POS_x(D))}{card(POS_C(D))}$ indicates the distinction ability of attribute X.

4). The basic operations of genetic algorithms

Take the tournament selection operator to select. And take the uniform crossover operator to crossover. Uniform crossover is that the gene is in each gene locus of two paired individual at the same crossover probability to exchange, so as to form two new individuals. Furthermore, take a basic bit mutation in the variations to mutate.

5). The mechanisms to prevent incest

In order to maintain the diversity of the populations when selecting the paired individuals, in this paper, we use the mechanism to prevent incest, restrict similar individual to mate. That is to say, according to the probability of selecting two individuals, if the Hamming distance between two individuals is less than the given threshold, then cross the two individuals in populations C; otherwise, get back and fetch them again.

6). The elitist strategy

In order to preserve the best individual of the fitness function value in the ancient individuals, in this paper, we take the elite strategy, that is to say, we copy the individual of highest fitness value in contemporary populations to the next generation population, and the individuals do not to participate in the operations of crossover and mutation.

2.3 The k-Means Clustering Algorithm

The classic k-Means algorithm is an unsupervised learning algorithm which is pre-determined and whose number of clusters is k. The clustering results are represented by. Suppose the sample set be $X=\{x_1,x_2,...,x_n\}$ and k-cluster centers be $c_1,c_2,...,c_k$. Let $w_1,w_2,...,w_k$ be k-cluster. Each cluster centre is $c_j=\frac{1}{N_j}\sum_{x\in w_j}x$, $j=1,2,\cdots,k$, where N_j denotes the j^{th} cluster's quantity. Define the criterion function as follows: $J=\sum_{i=1}^{k}\sum_{x\in w_j}d(x-c_i)$, where $d(x-c_i)$ denotes the Euclidean distance from the object of each cluster to the cluster centre. The processes of k-Means algorithm are as follows: at first, randomly select k objects, each of the remaining objects, according to their each of the distance from centre, it would be assigned to the nearest cluster, then calculate the average value for each cluster. This process is repeated until the convergence of the criterion function.

3. Efficient Clustering Method Based on Rough Set and Genetic Algorithm

Input: n numbers data objects

Output: Output the corresponding class centre and members of the object with the largest fitness function value.

Step 1: Coding k numbers clusters. The length of the chromosome is k+2, the first k numbers are the categories of the bit binary form, the k+1th number is the category of the decimal number, and the bottom number is the fitness function of the individual. Then obtain the corresponding k numbers categories of each chromosome, select k numbers objects as the cluster centre from the candidate point set according to the maximum-minimum distance principle, then according to the rough k-means clustering criteria to allocate each point.

Step 2: Initialize the population: randomly generate P numbers population chromosomes.

Step 3: Rough clustering: decode each chromosome in the group, then obtain the number of categories corresponding to each chromosome k, from the candidate point set in accordance with the principle of maximum and minimum distance of k objects selected as cluster centres, then follow the rough guidelines for distribution of k-means clustering points.

Step 4: To calculate the fitness function value: the fitness function is:

$$F(v)=\beta\cdot f(x)+p(x)=\beta\times(1-\frac{card(x)}{card(C)})+\frac{card(POS_X(D))}{card(POS_C(D))}$$

Step 5: Selection, crossover and mutation.

Step 6: If the number of iteration is equal to the defined maximum value, turn to step 7, otherwise, the algorithm continues from Step 3 to Step 5.

Step 7: The algorithm terminates.

4. Experimental Analysis

In order to verify the validity of the clustering algorithm, in this paper, we use five data sets in UCI machine learning database to test this algorithm. The data characteristic of the data sets is described in Table 1. Because the dates in data set have the determined classification, so the performance of the clustering algorithms can be expressed by accuracy. The experimental parameters are as follows: the population size M takes 30; the algorithm runs 100 times repeatedly, each time the initial populations randomly generate; take the way of roulette wheel selection to choose. From the results of Table 2, the experimental results show that the new algorithm effectively combines rough set and genetic algorithm,

which makes the clustering accuracy of the new algorithm superior to the accuracy of K-means clustering based on genetic algorithm[4].

Table 1. Experiential Dataset

Data Sets	No. of Instances	No. of attributes	No. of Classes
Soybean	47	35	4
Zoo	101	16	7
Dermatology	358	34	5
Pima	768	8	2
Mushroom	8124	22	2

Table 2. Comparison Results of K-means and Algorithm 1

Data Sets	No. of Classes	The Accuracy of K-means[4]	The Accuracy of Algorithm 1

Soybean	4	81.05%	85.32%
Zoo	7	78.40%	81.16%
Dermatology	5	80.29%	83.10%
Pima	2	82.14%	85.77%
Mushroom	2	83.50%	87.29%

5. Conclusion

Clustering is an unsupervised classification method, and it is without prior knowledge available in the data set. The traditional k-means algorithm and genetic algorithms need to determine the number of clusters and need to select the number of initial population by parameters. Furthermore, the improved genetic algorithm makes the results of heuristic clustering method not fall into the local optimum, which has strong global search ability. At the same time, the uncertainty of the cluster boundary object is represented by making use of rough set. So that the upper approximation and lower approximation sets in clusters can better describe the objective world. On this basis, an efficient clustering method based rough set and genetic algorithm is provided. The experimental results show that the algorithm can converge on the global optimal solution and can obtain better clustering results.

References

[1] Jiawei Han, Micheline Kamber. Data Mining: *Concepts and Techniques*[M]. US Kaufmann Publishers, Inc, 2001: p.223-262.

[2] Grabmeier J, Rudolph A. Techniques of cluster algorithms in data mining[J]. *Data Mining and Knowledge Discovery*, 2005, 6(4):303-360.

[3] Pawan Lingras. Interval Set Clustering of Web Users with Rough K-Means[J]. *Journal of Intelligent Information System*, 2004, 23: 15-16.

[4] Ting Lin, Haixiang Guo, Kejun Zhu, Siwei Gao. An Improved Genetic k-means Algorithm for Optimal Clustering[J]. *Mathematic in Practice and Theory*, 2007, 37(8):104-111.

[5] Pawlak Z. Rough set theory and its application to data analysis[J]. *Cybernetics and Systems*, 1998, 9: 661-668.

[6] Guoyin Wang, Yiyu Yao, Hong Yu. A Survey on Rough Set Theory and Applications[J], *Chinese Journal of Computers,* 2009. 32(7):1229-1246.