Research on the Novel Weighted Fuzzy Clustering Algorithm based on Fuzzy Sets and Rough Set Theory

Liwei Chen[1]

[1]College of Computer Science and Technology,

SouthWest University of Science and Technology

Mianyang,621010,China

Abstract. In this paper, we conduct research on the novel weighted fuzzy clustering algorithm based on fuzzy sets and rough set theory. Due to the large scale of data, in order to improve the efficiency of the clustering, we can use the attribute selection and data sampling to reduce the data size. We combine the characteristics of the fuzzy sets and rough set theory to optimize the prior objective function. In the near future, we plan to conduct more related research to polish the method.

Keywords: Weighted Fuzzy Clustering; Fuzzy Sets; Rough Set; Literature Review.

INTRODUCTION

With the rapid development of database technology and the wide application of database management system, more and more people to accumulate data. Hidden behind the explosion of data a number of important information, people want to be able to a higher level of analysis, in order to better use of these data. As an important branch of data mining, clustering analysis has attracted widespread attention, it can serve as an independent data mining tools or act as a preprocessing step for other data mining algorithm. Clustering is an unsupervised classification, is a kind of important people know the society and nature. On high-dimensional large spatial database, effective clustering algorithm to meet the six factors is necessary, however, the existing algorithm, although some algorithm can meet the requirements of some of these, even some can meet most of these factors, but almost no one can meet all the conditions. Clustering is the link between a given data set according to the object of the metrics are divided into several subsets, the process of making division within a subset of the class after high similarity, similarity between classes is low. Clustering by using the method of mathematical research and deal with the given data set. Clustering and classification is very different, in the classification, know in advance that a given data set contains several classes, in the process of classification and each data object which marked a class can be classified as. In the cluster, however, didn't know the number of clustering, at first through a standard will all data objects of different classes, eventually meet in the same class data similarity is the largest, not least the similarity of their kind [1-4].

Due to the large scale of data, in order to improve the efficiency of the clustering, we can use the attribute selection and data sampling to reduce the data size. Attribute choice is to remove those practical significance is small or no actual significance of attributes, or attribute set, reduce the dimensions of the data set. Data sampling is to use the principle of sampling, select a representative sample to represent the overall. Through the selection of attributes and data sampling have reduced the size of the data, processed data set is used to instead of the original data set, to enhance the validity of the data, reduced the space and time complexity of the algorithm, improve the efficiency of data

mining. In recent years, with the clustering algorithm is proposed and improved continuously, according to the different classification methods can be classified to different clustering algorithms, such as performance in different ways, according to the results of the clustering can be divided into hard clustering algorithm and fuzzy clustering algorithm and clustering algorithm. For hard clustering algorithm, a data object with the nature of the either/or, the data object can only be classified as a class of the class, due to a data object under different conditions can be classified into different classes, so hard clustering algorithm has significant limitations, and it's easy to fall into local optimal value.

According to the popular criteria, we could separate the clustering algorithms into the following parts. (1) Hierarchical clustering method. Hierarchical clustering method is composed of different levels of clustering segmentation, segmentation with nested relation between levels. Hierarchical clustering method of data collection of objects of a given level of decomposition, how to form, according to the level of decomposition can be divided into coherent and the division of the two methods. (2) Density clustering method. Density clustering method is to use data density function clustering. Its main idea is: as long as the density of the adjacent area of more than a given value keep clustering, and that is to say, for a given class within a given range of each data point in the region must contain at least a certain number of points. (3) Grid clustering method. Grid clustering method is quantitative data object space into a finite number of units that form a grid structure. All cluster operations are conducted on the grid structure. (4) The method based on constraint. Clustering technology for the development of practical applications, such as geographic data analysis provides many useful tool, but the vast majority of clustering algorithms cannot be directly solve reality constrained clustering problems.

To enhance the current method, we conduct research on the novel weighted fuzzy clustering algorithm based on fuzzy sets and rough set theory. Researchers has put forward many improved methods, have a plenty of based on the idea of weighted, have a plenty of based on genetic algorithm and artificial immune clone selection and the biological theory of technical thought, also have a plenty of based on possibility theory. In the figure one, we illustrate the conceptual demonstration of data and feature clustering. In the following sections, we will discuss the theory in detail.

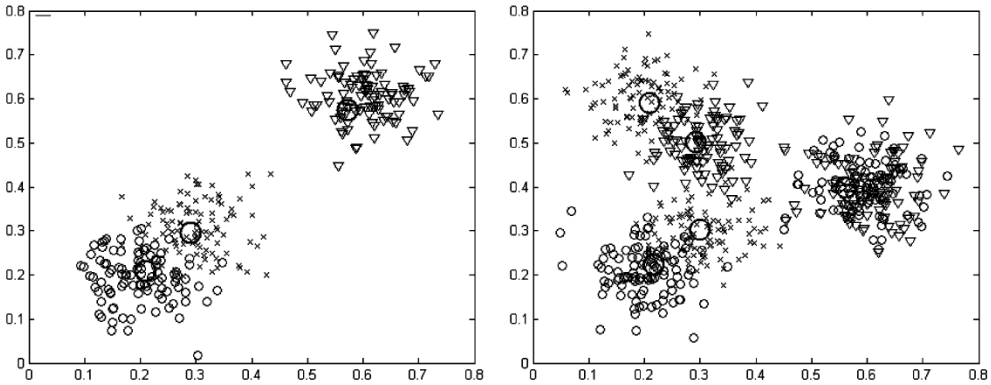

Fig. 1The Conceptual Demonstration of Data and Feature Clustering

The Proposed Algorithm

The Rough Set Theory. Rough set theory is defined to be the initial prototype from the relatively simple information model, its basic idea is formed by relational database classified concepts and rules, through the classification of the equivalence relation and classification for the target approximate knowledge discovery. At the core of rough set theory and application of approximate space is derived from a pair of approximation operators, namely approximate operator and lower approximate operator. No clear relation in the classical model is a kind of equivalence relation, demanding, limits the application of rough set model. Therefore, how to promote define approximate operator has become a key point in the research of rough set theory. The set is defined as the follows.

$$R(X)=\left[a\in U:[a]_R\subseteq X\right]$$

(1)

In fact, there are two kinds of forms to describe rough set, a set from the point of view and one is from operator's point of view. So, from a different point of view using different research methods are various extensions of rough set model. Extension model of research and applications based on its research has become a new research hotspot. The formula two defines this.

$$BN(X)=R^{*}(X)-R_{*}(X)$$

(2)

Rough sets theory and other theories of dealing with uncertainty and imprecise problems is the most significant difference about processing for it does not provide any priori information data collection, so the description of the uncertainty of a problem or processing can be said to be more objective. Based on the application of rough set theory research mainly concentrated in attribute reduction, rule acquisition, computational intelligence algorithm based on rough set research.

$$H(P)=-\sum_{i=1}^{m}p(X_i)\log p(X_i)$$

(3)

The basic thought of classical rough set theory is based on the equivalence relation of graining and the approximate method of data analysis. At the core of rough set theory and the general application of approximate space is derived from a pair of approximation operator, namely approximate operator and lower approximate operator. And then to find a binary relation that generated by the binary relation and its methods of approximate space according to the structure of export of approximation operator is just given by the axiomatic method definition set operator. Some special axiom of the approximation operators to ensure that there are some special types of binary relation exists, enables these relationships by constructing methods produce a given operator. On the other hand, by the binary relation derived by constructing method of approximation operator must satisfy certain axioms, make these axioms by algebraic method to produce a given binary relation.

The Weighted Fuzzy Clustering Algorithm. Based on the hierarchy clustering can be divided into two methods: bottom-up method and split the top-down method. Condensation methods refers to each data object as a class, and then according to the similarity between data gathered into a larger class, until all objects of a class or meet the conditions to stop the clustering. Anti-secession law is just the reverse of condensation method implementation process, the first will be all of the data object as a class, and then according to the data between the thin subdivided, until every data objects into a class or meet the conditions to stop clustering. Because based on the structure of the objective function to use clustering prototype, so the structure is based on the objective function of clustering to determine the clustering prototype. In the process of the development of clustering, clustering prototype research, in the beginning of clustering analysis in order to measure the super sphere

structure, so the center node in space for the data points, namely, data center, if the clustering structure is super sphere, should choose other data. The objective function is shown in the formula four.

$$F(m)=\{J_m:1\le m\le \chi\}$$

(4)

Reality often face data contains some logo, have noise, the combination of a variety of data types, data clustering algorithm is put forward by the based on the proposed on the basis of different data types. Optimization objective function method is essentially the process of iterative optimization, so it will often get the local optimal value rather than the global optimal value, and is especially important in the initial stages of clustering and it directly affects the later clustering effect. Extensive research algorithm is more sensitive to change the initialization for faults, only true to overcome this shortcoming, can obtain the stable clustering results. Applying neural network to data clustering, mainly considering the obvious advantages of neural network, it can parallel processing data, the advantage is more obvious when particularly large data sets, however, the aforementioned neural network is not very perfect, it will not be able to discover clusters of arbitrary shape, for clustering, is a very important problem. The realization of the evolutionary computation is based on and develops on the basis of biological evolution, because this method can realize parallel global search, thus to obtain the global optimal solution of the possibility is very large, and this method is easy operation, universality, has the advantages of good processing capability of noise data, therefore, on the basis of the advantages of this method and it is introduced into the fuzzy clustering.

$$v_i^{(T+1)}=v_i^{(T)}+\lambda\left(x_k-v_i^{(T)}\right)$$

(5)

Due to the fuzzy c partition can draw between each sample and each class membership degree, scope and different external list to differentiate between cohesion and discrete condition, thus can get more information better.

The Numerical Analysis and the Simulation. To conduct the thorough research to the general perturbation of the fuzzy clustering algorithm, give an example of fuzzy similar matrices there will have the same distance with the matrix and the smallest two unequal fuzzy equivalence matrices which proves that the global optimal fuzzy equivalence matrix has uniqueness. Based on fuzzy clustering algorithm shortcoming many improved algorithm to adapt to the different directions of applications, such as, based on the objective function based on similarity relation and fuzzy, based on the data set of convex decomposition of the relationship between the different improvements such as direction. By introducing a parameter to the weighted average of the error within the class and function of target function is extended to the infinite variety. One of the most popular for clustering method based on the target which converts the problem to nonlinear programming problem with constraints, thus to solve the data set we finalize the experiment in the following part. In the figure three, we illustrate the performance and simulation of the proposed methodology.

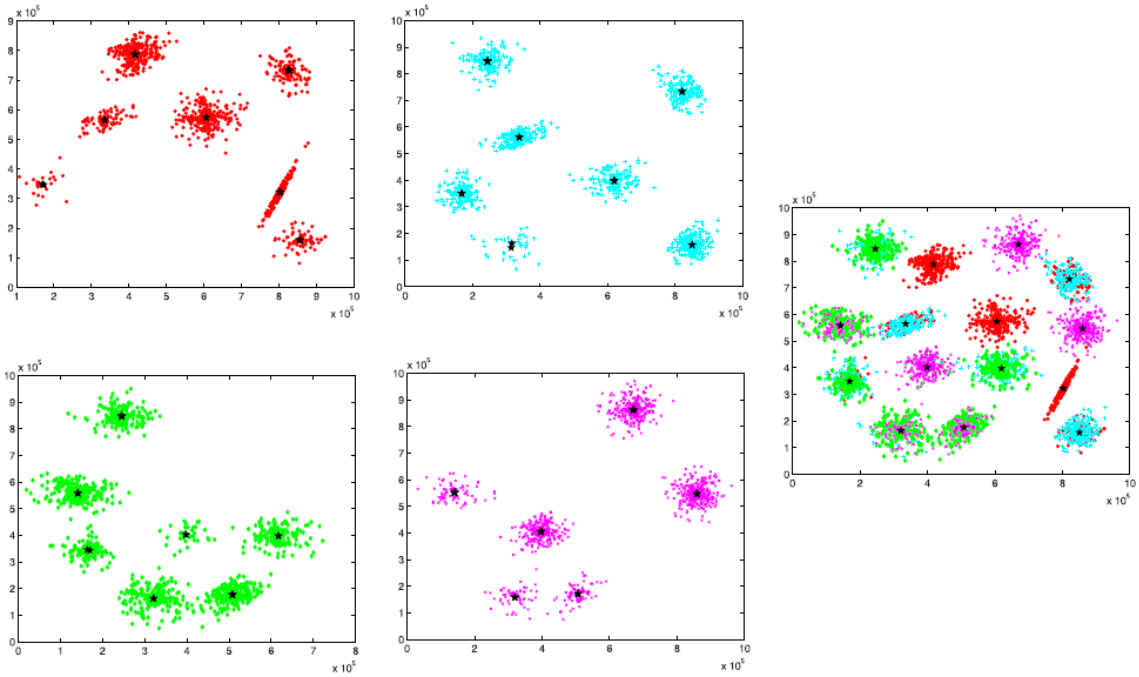

Fig. 2The Performance and Simulation of the Proposed Methodology

CONCLUSIONS

In this paper, we conduct research on the novel weighted fuzzy clustering algorithm based on fuzzy sets and rough set theory. Clustering is an unsupervised classification which is a kind of important method people knows the society and nature. Our experimental result and simulation shows that the designed algorithm outperforms the other methods which will enhance the current research on the topic of machine learning and artificial intelligence.

Acknowledgement

This research is financially supported by the artificial intelligence laboratory in Sichuan province, the key laboratory open fund (NO. 2014RYY03).

References

[1] Rahulamathavan Y, Veluru S, Phan R C W, et al. Privacy-preserving clinical decision support system using Gaussian kernel-based classification.[J]. Biomedical & Health Informatics IEEE.

[2] Tan X, Yu X, Qin J, et al. Multiple kernel SVM classification for hyperspectral images[J]. Chinese Journal of Scientific Instrument, 2014, 35(2):405-411.

[3] Wang H, Wang J. An Effective Image Representation Method Using Kernel Classification[C]// Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on. IEEE.

[4] Wang F, Zuo W, Zhang L, et al. A Kernel Classification Framework for Metric Learning[J]. IEEE Transactions on Neural Networks & Learning Systems, 2014.