Research and Application of High Dimensional Discrete Data Clustering Algorithm Based on Kernel Function

Fulan Ye

Fuzhou University of International Studies and Trade, Fuzhou, Fujian, 350202

Keywords: Research; Application; High Dimensional Discrete Data; Clustering Algorithm; Kernel Function

Abstract: Aiming at the disadvantages of traditional fuzzy kernel clustering algorithm, which do not consider the different contribution degree of each dimension feature to clustering, and easily fall into local optimum, an improved fuzzy kernel clustering algorithm is proposed. The algorithm constructs a simple and effective fitness function that combines the advantages of the global search of the genetic algorithm to avoid the algorithm falling into a local optimum. A weight coefficient was also introduced for each dimension feature and weighted with the Relief algorithm. This algorithm is much better than the traditional fuzzy kernel clustering algorithm. The experimental results show its effectiveness.

1. Introduction

Cluster analysis is an important branch of unsupervised pattern classification in statistical pattern recognition. It classifies a given set of unlabeled samples into multiple categories according to certain criteria, so that samples in the same class have higher similarity. , And the samples in different categories differ greatly. With the introduction of fuzzy theory, due to the ambiguity of the classification, people gradually accepted the fuzzy clustering analysis [1]. Among many implementations, fuzzy C-Means (FCM) has become one of popular algorithms. One of the biggest drawbacks of FCM is that the data clustering effect on the hypersphere structure is significant, but it is not effective on non-hypersphere structure data. The success of support vector machines has aroused people's interest in the study of nuclear methods. Literature has done pioneering work on combining nuclear methods and clustering algorithms, and proposed a hard-dividing kernel clustering algorithm. The literature [2] extended it to fuzzy C-means clustering and proposed a fuzzy kernel C-means clustering algorithm (Kemelized Fuzzy C-Means, KFCM).

The basic idea of KFCM is to map the data nonlinearly to the high-dimensional feature space through different kernel functions, so that the features that have not been shown originally emerge, expand the differences between features, and then perform fuzzy C-means aggregation in high-dimensional feature space. This to some extent overcomes the shortcomings of FCM not suitable for multiple data structures. However, KFCM still has the following two main disadvantages: 1) Because the Euclidean distance based on nuclear space is still used, the different contribution degree of each dimension feature to clustering is not considered. In the actual situation, some features play an important role in the clustering process. The role of some features may even be neglected. The contribution degree of each dimension feature to clustering is different. 2) It is easy to converge to the local optimal value and sensitive to the initial value due to the iterative solution using the gradient descent method. For this reason, for the former, we introduce the weight coefficient into the objective function and set a weight coefficient for each feature. For the latter, we use the characteristics of the global search of the genetic algorithm, encode the partition matrix, and design a simple and effective fitness function.

2. High-Dimensional Discrete Data Clustering Algorithm Based on Kernel Function

Weighted Fuzzy Kernel Clustering. KFCM maps the sample space to the high-dimensional feature space through Mercer kernel function nonlinearity, which amplifies the feature differences

between samples, thereby improving the clustering effect. However, KFCM is still based on the Euclidean distance of nuclear space. The Euclidean distance assumes that each feature has the same importance in the clustering process. In the actual situation, some features play an important role in the clustering process, and some features may even be ignored. In view of this, this paper introduces the feature weight coefficient into the objective function [3].

This paper uses the method of determining the weight coefficient matrix of the literature to weight the features using the ReliefF algorithm. The Relief algorithm randomly selects m sample instances from the training set. Through the difference between the selected sample and the two nearest neighbors belonging to the same class and different classes, the correlation between the characteristics of each sample and the class is calculated and then averaged. However, the Relief algorithm is limited to solving two types of classification problems. The ReliefF algorithm is an extended algorithm of Relief. Instead of selecting only one nearest neighbor sample from the same class and different classes, the ReliefF algorithm selects the k nearest neighbor samples and averages them to obtain each feature weight. The ReliefF algorithm can solve many problems. However, the ReliefF algorithm is aimed at the classification technique. In the classification, the category label of each sample is determined. In the cluster analysis, the category label of each sample is unknown. Therefore, the sample set to be analyzed is firstly processed once. Clustering, obtaining the class labels of each sample, and then using the ReliefF algorithm to calculate the degree of contribution of each dimension feature to the cluster. Finally, the result obtained by normalizing ReliefF is the matrix of weighted coefficients.

Fuzzy kernel clustering based on genetic algorithm. Since KFCM, like FCM, is essentially a local search algorithm based on gradient descent, it inevitably falls into a local optimum, and it is sensitive to the initial value. Genetic algorithm is an algorithm with strong global search ability that simulates natural selection and genetic variation in the process of biological evolution. Therefore, genetic algorithm is very suitable for solving clustering problems. The literature proposes a hybrid genetic algorithm that replaces crossover operators with K-Means operators. The literature adopts floating-point code clustering centers and designs floating-point crossover and mutation operators to improve search efficiency. However, they are prone to precocity when the number of samples, dimensions and categories are large. And just searching in the sample space, the clustering effect on non-hyperpheric data structures is not good. Therefore, this paper proposes a hybrid KFCM algorithm based on genetic algorithm. The design is mainly from two aspects: On the one hand, floating-point code partition matrix, and a simple and effective fitness function suitable for kernel clustering is designed. On the other hand, to improve the convergence speed, a KFCM optimization is added in each iteration.

In the genetic clustering algorithm, there are two encoding schemes: one is to encode the partition matrix U; the other is to encode the cluster center V. This article uses floating-point coding of the partition matrix, which is more convenient and easier to understand. In the floating-point coding method based on partition matrix, a chromosome consists of a fuzzy partition matrix: chromosome=U. Divide n samples into c classes, so a chromosome is a $c \times n$ matrix. After the encoding method is determined, the population is initialized. Assuming that the population size is N, N fuzzy division matrixes are randomly generated to constitute the initial population [4].

The selection operation uses a combination of an optimal preservation strategy and a roulette ratio selection method. 10% of the excellent individuals in the population go directly to the next generation, and the remaining populations are selected by roulette. In this way, the average fitness value of the group can be continuously improved, and the fitness value of the optimal individual cannot be reduced. Crossovers use single-point crossings, and mutations use single-point mutations.

Because the weighted KFCM has a strong local search capability, it performs a weighted KFCM optimization after performing each genetic operation, and the optimized group enters the next generation population to improve the convergence speed of the algorithm.

The success of support vector machines has rekindled people's interest in nuclear methods. The basic idea of nuclear methods is to use nonlinear transformations to map the input mode space R^p to a high-dimensional feature space R^q, and then to design linear learning in high-dimensional feature

spaces. If the interaction between each mode vector in the algorithm is limited to the inner product operation, there is no need to know the specific form of the nonlinear transformation, as long as the inner product in the linear algorithm is replaced by a kernel function that satisfies the Mercer condition, and the original input can be obtained. The corresponding non-linear algorithm in space, this method called kernel technique, uses the Mercer kernel function to simplify the calculation. It can solve the dimension disaster problem. For example, it represents a nonlinear mapping function. Commonly used kernel functions satisfying Mercer conditions include polynomial kernel functions and radial basis kernel functions.

3. High-Dimensional Discrete Data Clustering Algorithm

Cluster analysis is widely used in fields such as data mining, computer vision, and unsupervised pattern recognition. The main task of clustering is to divide given data into c(c>1) subsets (clusters), making them in the same subset. There is a high degree of similarity between samples in (clusters) and between samples in different subsets. Compared to traditional hard clustering algorithms, fuzzy clustering algorithms have more Good data expression ability and clustering performance. Fuzzy c-means clustering algorithm (FCM) is one of the most widely used fuzzy clustering methods. FCM is simple in calculation and good in clustering, but it is very sensitive to noise and outliers. Recently, Pal et al. [5] proposed the ambiguity c-means clustering algorithm (PFCM), which has been widely used due to its better solution to noise-sensitivity problems in FCM. However, there are many parameters in the PFCM algorithm. Usually need to be manually specified, lack of theoretical basis, so that the algorithm has a strong dependency on the parameter settings. Ng and Jordan et al. proposed the Ng-Jordan spectral clustering method, which is also an important clustering algorithm that is widely used at present. It can find differences the clustering structure of the shape, but the clustering result is sensitive to the selection of the similarity matrix, the complexity is high, and it is not suitable for large data sets. Although the sparse similarity matrix can be chosen to partially overcome this problem, how to construct a reasonable and effective similarity matrix is still a problem to be solved. The nuclear method has been applied to many aspects of machine learning, such as nuclear principal component analysis, nuclear F isher Discriminant analysis and kernel-based clustering analysis. Clustering based on kernel method maps the points in the original space to the feature space through kernel functions. The algorithm design, analysis, and calculation are performed directly or indirectly in the feature space to obtain the original spatial clustering. In this paper, a new clustering algorithm IKFCM is proposed, which uses nuclear techniques combined with improved fuzzy c-means algorithm clustering rules. It can overcome the noise in PFCM to a certain extent while overcoming the PFCM. Sensitivity of parameter setting, the algorithm has lower spatial complexity, and is more suitable for clustering large datasets than the Ng-Jordan method. Comparison of artificial and real data with various algorithms shows that the IKFCM algorithm is relatively low Time, space complexity and more accurate clustering results.

4. Experimental Results

In order to test the performance of this algorithm, the famous IRIS actual data set was used as the test sample set. IRIS datasets are often used as standard data to test the performance of an algorithm. It consists of 150 sample points in a four-dimensional space. We use classic FCM, traditional KFCM and the algorithm of this paper to classify IRIS samples and test the number of misclassified samples of the three algorithms to compare their performance. The traditional FCM algorithm misclassified 16 samples, and the traditional KFCM misclassified 10 samples. The algorithm of this paper only misclassifies 6 samples. Experimental results show that the performance of this algorithm is greatly improved.

5. Conclusion

In summary, the paper uses the kernel function to map the sample points to the high-dimensional

feature space, and uses the weighted error square sum criterion as the clustering basis for clustering in the feature space. The Gaussian kernel function is used for comparison experiments. The results show that the proposed algorithm is more robust than the PFCM algorithm. The membership value obtained can more reasonably reflect the degree of similarity between each sample point and each cluster, and at the same time overcome the PFCM algorithm due to the introduction. There are many parameters, and the clustering result is sensitive to the parameter setting. Compared with the Ng-Jordan method, the given algorithm has faster speed and lower space complexity.

Acknowledgements

Fund Project: 2015 Fujian Young Teacher Education Research Project (Science and Technology): Research and Application of High-Dimensional Discrete Data Clustering Algorithm Based on Kernel Function Project Number: JA 15586, Project Leader: Ye Fulan

References

[1] ATANASSOV K. Intuitionistic fuzzy sets [J]. Fuzzy Sets and Systems, 2006, 20(1): 87 - 96.

[2] Zhang Hongmei, Xu Zeshui, Chen Qi. Intuitionistic fuzzy set clustering method [J]. Control and Decision, 2007, 22(8): 882-888.

[3] Shen Xiaoyong, Lei Yingjie, Li Jin. Intuitionistic fuzzy clustering method based on objective function [J]. Systems Engineering and Electronics Technology, 2009, 31(11): 2732 - 2735.

[4] Shen Xiaoyong, Lei Yingjie, Cai Ru. An Intuitionistic Fuzzy Clustering Initialization Method Based on Density Function [J]. Computer Science, 2009, 36(5): 197-199.

[5] Xu Xiaolai, Lei Yingjie, Zhao Xuejun. Intuitionistic Fuzzy Clustering Based on Intuitionistic Fuzzy Entropy [J]. Chinese Journal of Air Force Engineering, 2008, 9(2): 80 - 83