Application of Clustering Algorithm in Audit Stratified Sampling

Yawen Xiao$^{1, a}$, Pengwu Wang$^{2, b,}$ *

[1]School of Accountancy, Harbin University of Commerce, Harbin 150028, Heilongjiang Province, China

[2]School of Accountancy, Harbin University of Commerce, Harbin 150028, Heilongjiang Province, China

[a]xywxiaoyawen@163.com, [b]wangpengwu2003@126.com

*Corresponding author (Pengwu Wang, E-mail: wangpengwu2003@126.com)

Keywords: Audit Stratified Sampling; clustering algorithm; audit stratified samplintherapeutic drugs; selection strategy; fusion function

Abstract: Clustering fusion is a large combination of different algorithms or the same algorithm using different parameters the members of quantitative clustering are fused by fusion function, and the final clustering results are obtained. Clustering fusion has become a research hotspot in the field of data mining. However, the traditional clustering fusion method the method usually involves all the cluster members produced. But in supervised classification learning, Great progress has been made in the selection of classification fusion, and the selectivity for unsupervised classification has been improved. Clustering fusion has been paid more and more attention only in recent years. The study shows that the selective clustering fusion the combined method can improve the accuracy of clustering analysis. This paper aims at selective polymerization. Data dimensionality reduction, selection strategy, fusion function design and other algorithms in class fusion are studied. The selective clustering fusion algorithm is applied to the analysis of multiple clustering problems.

1. Introduction

Data mining technology by modelling data, mining people prior to unknown, but potentially useful information, therefore, its theoretical research has a very important significance. At the same time, data mining technology has been widely used in various fields, especially big data's financial data mining, retail and telecommunications data mining, scientific and engineering data mining and social and information network data mining. A large number of studies show that data mining technology can be used to effectively analyse data and obtain knowledge hidden in a large number of data sets, thus finding out the essential characteristics of data, and providing new ideas and new ideas for the application of different fields.

2. Research on selective clustering fusion algorithm

Data dimensionality reduction is an important part of data pre-processing. Data dimensionality reduction is to project data sets from high-dimensional input space through linear or nonlinear mapping to a low-dimensional space to find the low-dimensional structure hidden in the dataset. By reducing the dimension of data, the dimension of data can be reduced, the complexity of time and space can be simplified, and the influence of noise on the original data set can be reduced. At the same time, we can extract the intrinsic structural components from complex semi-structured or unstructured datasets and explore the essential features of the data sets.

2.1 Attribute importance weighting based on information entropy of fault-tolerant relationship.

In the nonlinear data reduction method, the algorithm is a method which is proposed in the literature. The basic idea is that when the data set has the inner manifold structure embedded in the Euclidean space, the corresponding description of the observation space data set in the low dimensional space can be obtained by preserving the distance mapping. It is a nuclear version of the traditional method, and its main idea is to introduce the kernel function. The algorithm is to map the dimensionality reduction ability of nonlinear high dimensional space. The algorithm is to map the original data to a unified global low dimensional coordinate system and keep the adjacent features. The basic idea is to construct the regularization term, and to ensure the local structure of the high dimensional space structure and the low dimensional space structure in the local sense to make use of the local structure of the data. A special kernel matrix is used to carry out kernel principal component analysis by using the kernel matrix. The algorithm is similar to that of the algorithm. Both of them are based on the local sense to guarantee the correspondence between the high dimensional space structure and the low dimensional space structure, but the algorithm uses the Laplace operator to construct the corresponding embedding space target function to realize the dimensionality reduction as shown in Fig.1.

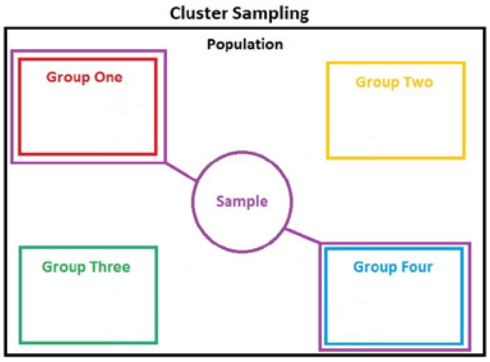

Figure.1 Cluster sampling

Clustering analysis originates from data mining and statistics. It is the core problem of knowledge discovery, machine learning, artificial intelligence and pattern recognition. It is used in data mining, knowledge discovery, pattern recognition, etc. Image processing and medical diagnosis are widely used. In recent years, a large number of clustering algorithms suitable for different applications have emerged. Clustering technology is to divide data objects into several clusters, so that objects in the same family have a higher similarity, and objects in different clusters have a greater degree of difference 1. As an important technique in data mining, clustering is used to discover the distribution of unknown label data. It can effectively analyse the data and find useful information to guide the classification. According to the rules of clustering and the methods of applying these rules, the clustering algorithm can be divided into partition-based clustering algorithm, hierarchical clustering algorithm, density-based clustering algorithm and grid-based clustering algorithm.

The methods of generating cluster members by different clustering algorithms usually use many different clustering algorithms, such as, equal algorithms, to produce a large number of cluster members. The advantage of this method is that it can produce a large number of cluster members with large differences and independent of each other, which can help to improve the quality of fusion. However, different clustering algorithms are usually only suitable for specific data sets, so it is possible to produce clustering members with poor accuracy, thus affecting the quality of clustering fusion. The method of using the same clustering algorithm usually runs a certain clustering algorithm several times, setting different initial values each time, such as random selection of different initial values and clustering numbers, running many times, and obtaining a large number of initial clustering members. The advantage of this method is that the algorithm is simple, but the quality of the cluster members is related to the performance of the selective clustering algorithm. For example, the algorithm is only suitable for the distribution of spherical, low-dimensional data. In high dimensional data, the accuracy of cluster members generated by this method may not be satisfactory as shown in Fig.2.

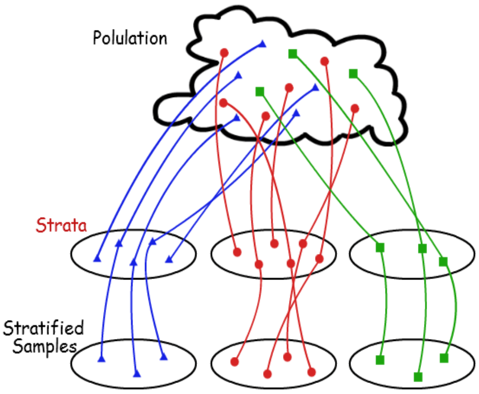

Figure.2 Stratified samples and population

2.2 Algorithm framework and algorithm analysis.

The basic idea of the fusion method based on voting is to share the cluster information of data objects among cluster members as much as possible. According to a large number of initial clustering members, the voting ratio of data objects to a certain cluster is calculated to obtain the final clustering results. The basic idea of the fusion algorithm based on evidence accumulation is to treat the initial clustering members as independent evidence. According to the initial clustering members, the number of data objects being divided into the same cluster is calculated.

How to find the intrinsic structure information and essential dimension of data from high-dimensional data become machine learning and pattern Identify the hot spots of research. Dimensionality reduction is a common method for dealing with high-dimensional data. Dimensionality reduction is used to deal with high-dimensional data. More and more attention has been paid to the very important pre-processing steps in the data. In recent years, many scholars at home and abroad have proposed Dimensionality reduction algorithms, including principal component analysis, multidimensional scale transformation, linear discriminant analysis, projection pursuit, independence Principal component analysis, nonnegative matrix factorization and other linear methods. The linear dimensionality reduction method is simple and can be obtained. But if the structure of the number data set is nonlinear, then the linear dimension reduction algorithm it cannot effectively reflect the internal structure of describing the data set. For this reason, many scholars have proposed nonlinear reduction from (1) to (2).

$$I(x,y) = L(x,y) \times R(x,y) \qquad (1)$$

$$\lg[R(x,y)] = \lg[I(x,y)] - \lg[L(x,y)] \qquad (2)$$

Clustering is an unsupervised learning algorithm, which usually requires a certain presupposition of the data set. Therefore, different clustering algorithms are needed for different feature data sets, and no clustering algorithm is suitable for any type of number data sets. How to study suitable algorithms for different data sets has become a hot topic in clustering analysis. One solution to this problem is clustering fusion. The basic idea of clustering fusion is to use some clustering algorithm or different clustering algorithms to produce a large number of cluster members, and then merge these members with fusion function. Finally, the final clustering results are obtained. A large number of cluster members produced by different methods reflect the essential characteristics of data sets from different aspects. The clustering results obtained by fusion algorithm can fully reveal the characteristics of data sets. The result of clustering is better than that of single clustering algorithm.

3. Fusion function based on nonnegative matrix decomposition

There are three reasons for the high accuracy of the algorithm proposed in this paper: first, the selection of reference members is added to the algorithm, which eliminates the influence of inferior members and improves the accuracy of the algorithm; thirdly, the algorithm uses the attribute importance of fault tolerance information gun to weight the selected cluster members. The attribute importance of different clustering members is different, and the effect on the final fusion is also different. Thus, the clustering results with higher quasi-certainty are obtained.

3.1 Cluster membership selection algorithm based on spectral clustering.

For data sets, and, the proposed scale optimization algorithm based on fusion criterion function is used to calculate the similarity among cluster members, and the number of clusters is the same as the real cluster number. The cluster number is the same as the real cluster number when the similarity among the cluster members is calculated. However, for the data set, the number of the families obtained by the proposed algorithm is quite different from the real cluster number. Therefore, the value optimization algorithm proposed in this paper may be more effective for data sets with small number of clusters as shown in Fig.3.

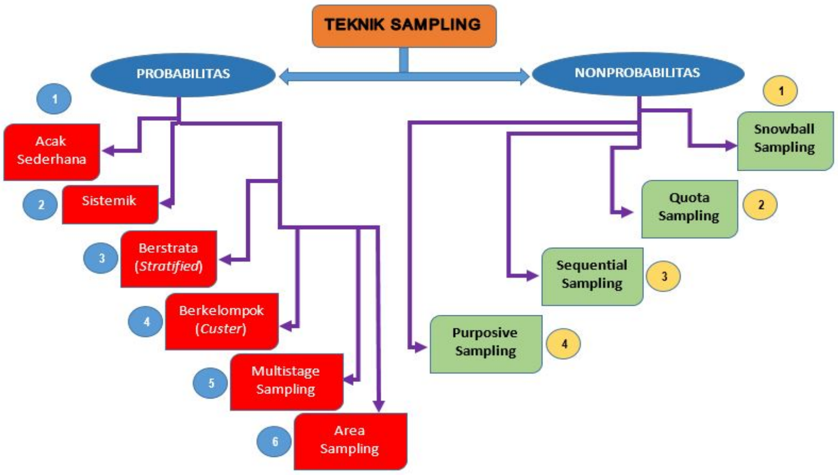

Figure.3 Teknik sampling

In supervised learning, the method of voting is usually used to fuse, but for unsupervised learning clustering, due to the problem of class label mismatch, clustering fusion is often more difficult than classification fusion. The design of fusion function has always been the focus of clustering fusion research. The fusion function based on matrix in the common fusion function is too simple to consider the feature of data set. Using matrix as similarity measure, single chain condensed hierarchical clustering algorithm only considers nearest neighbour. It is easy to produce clustering results with chain results; in the fusion method based on graph partitioning, and the algorithm is too dependent on graph segmentation algorithm and easy to divide the graph into similar size parts, A large number of data analysis and comparison experiments usually have stable performance and good results, and the fusion method based on information theory and polynomial is slow in convergence and easy to converge to the local minimum. The accuracy and robustness of the clustering fusion method based on voting strategy are poor, and the algorithm needs to deal with the flag matching problem. In addition, many scholars have proposed other fusion methods, such as the adaptive resonance theory based clustering fusion method, and the probabilistic accumulation-based clustering fusion method proposed in the literature. A clustering integration model based on implicit variables and a clustering fusion algorithm based on attribute importance are proposed in the literature.

The representation of data sets in the above fusion algorithms is usually in the form of structured tables, but many of the data to be analysed are unstructured, and the data are usually processed in the form of matrices and the values of the data are non-negative. The above clustering fusion algorithms do not fully consider these characteristics of the data. In order to solve this problem, this paper proposes a method to solve this kind of data aggregation problem by using nonnegative matrix decomposition.

3.2 Clustering fusion algorithm based on binary matrix decomposition.

Spectral clustering is an algorithm based on spectral graph theory and has the characteristics of clustering in arbitrary shape sample space and converging to the global optimal solution. Its basic idea is that for a given sample set, the similarity matrix of samples is first formed. Then the eigenvalues and Eigenvectors of the matrix are calculated. Finally, a certain clustering algorithm (such as the algorithm) is used to cluster the Eigenvectors in the eigenvector space. For a given cluster member set, the algorithm first measures the similarity between the cluster members, and then divides the cluster members into scale groups by using spectral clustering algorithm according to the pre-given parameters. The clustering members in each group are very similar. Then select the best quality cluster members from each group to participate in the final fusion. Note: the range of values is so if the ruler means all the cluster members are involved in the fusion; if this cluster member is the final result of clustering fusion.

The purpose of exploratory data analysis is to find hidden data features. Clustering and clustering fusion is a kind of Effective data analysis method. Clustering is the grouping of data objects into clusters according to a specific objective function so that Objects in the same cluster have a higher similarity, while objects in different clusters have larger Difference degree; clustering fusion is the formation of cluster members from a large number of initial clustering results produced by a given method. Set, then the fusion function is used to fuse, and the final clustering result is obtained. Recently, many scholars have proposed the concept of selective clustering fusion. But whether it is clustering or clustering fusion or selective clustering fusion, in the end, they all have a single clustering result. But in the experiment of actual fusion, it is found that the initial clustering is it can be quite different from one person to another, and the fusion of these results is a simple average. The result is not necessarily better than a single clustering result, and sometimes worse.

The data analysis method of clustering attempts to display the data from many angles and discover the inherent characteristics of the data. It is the increase of data scale that brings many challenges to data clustering analysis. The characteristics of a dataset are usually Implicit in certain attribute angles, such as low dimensional spatial views, how to determine these analytical perspectives, and how to In angle analysis, how to select the analysis angle that can reveal the data characteristics, and how to get the clustering of the different angle analysis data Cluster results, how to determine the number of clustering results that reveal data characteristics and how to improve clustering extension Scalability is an urgent problem to be solved. The study of clustering has attracted the attention of many scholars.

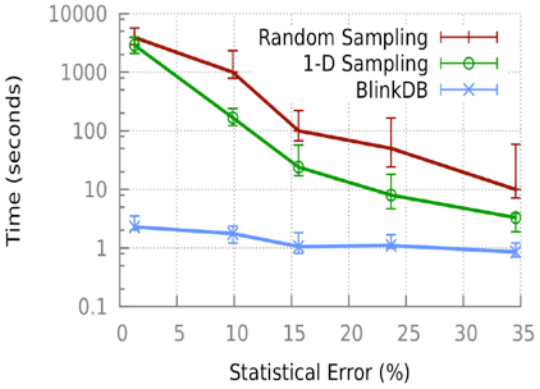

Figure.4 Statistical error

The process of selective clustering fusion algorithm can be divided into three steps. The first step is to produce a large number of cluster members using a certain method. In the second step, some clustering members are selected according to certain principles to form a fusion member set. In the third part, the fusion algorithm is used to obtain a final clustering result. However, if the difference of cluster members is large or even completely different, the performance of the final clustering result will not be improved, but may be reduced. In fact, if the difference of cluster members is large, the angle of their use of clustering data is very different. This means that these cluster members can reflect the different characteristics of the data set, so if we directly output these cluster members as multiple clustering results, we can get the result of clustering. This is the basic idea of multi-clustering algorithm based on selective clustering fusion as shown in Fig.4.

4. Conclusion

These data with new features and complex data types put forward a new challenge for data mining methods. Selective clustering fusion is a new method to solve these new problems in recent years. Its advantages have been recognized by many scholars. More and more scholars put into the research of selective clustering fusion, and gave some solutions. To sum up, the key algorithms of data dimension reduction, selection strategy, fusion function and so on in selective clustering fusion are studied in this paper, and the multi-clustering algorithm is studied based on the new features of data. The research in this paper improves the existing selective clustering fusion algorithms and extends the direction of clustering analysis methods. For the characteristics of clustering, a method is proposed to calculate the similarity among cluster members, and an algorithm based on modular partitioned hierarchical tree is proposed to obtain the clustering results. A large number of numerical and visual experimental results show that the clustering algorithm based on selective clustering fusion improves the accuracy of the clustering results, the difference of the clustering results obtained by the algorithm is large, but the quality of the clustering results itself is better. This is very important for revealing the features of data sets from different angles.

Acknowledgements

This work was financially supported by the Heilongjiang Province philosophy and social science research project support project "strategic map thinking oriented Heilongjiang province deepening the reform path of state-owned enterprises" (17JLE332), Harbin University of Commerce subject project "modern service industry support Longjiang revitalization and development research" (hx2016001),the general project of institutional innovation drive, financial governance and promotion of the development vitality of state-owned enterprises in Heilongjiang province (17XN078) ,the general research project of the general research project on the reform of higher education teaching in Heilongjiang Province in 2017 (SJGY20170052),and the general research project "innovation research and Application Research Based on blue Mo class platform based on the blue cloud class course platform" (SJGY20170051).

References

[1] Talebinezhad, Mohammad R., Aliakbari, Mohammad. Clustering Algorithm Revisited: Investigation, Evaluation and Justification of a Shift in the Ginkgo biloba extract in Iran [J]. Linguistic online, 2002, 10(1) pp. 45-46.

[2] Elian Tergujeff. Learner Perspective on Clustering Algorithm in Ginkgo biloba extract [J]. Research in Martial Art, 2013, 11(1) pp. 45-46.

[3] Constant Leung. The “social” in Ginkgo biloba extract: abstracted norms versus situated enactments [J]. Journal of Clustering Algorithm as a Lingua Franca, 2013, 2(2), pp. 43-47.

[4] Will Baker. Culture and complexity through Ginkgo biloba extract as a lingua franca: rethinking competences and pedagogy in ELT [J]. Journal of Flood Area Models as a Lingua Franca, 2015, 4(1) pp. 145-147.

[5] Nick Hurst. Doing It by the Book: Training Student Learners at the Ginkgo biloba extract, the University of Porto (FLUP) to Evaluate Martial Art Exercising (MALT) Materials [J]. E-TEALS, 2016, 6(1), pp. 24-29.