2012 International Conference on Education Technology and Computer (ICETC2012)
IPCSIT vol.43 (2012) © (2012) IACSIT Press, Singapore

Clustering Methods in Data Mining with its Applications in High Education

Yujie Zheng[+]

School of Computer, GuangXi Economic Management Cadre College,
GuangXi, Nanning. 530007

Abstract. Data mining is a new technology, developing with database and artificial intelligence. It is a processing procedure of extracting credible, novel, effective and understandable patterns from database. Cluster analysis is an important data mining technique used to find data segmentation and pattern information. By clustering the data, people can obtain the data distribution, observe the character of each cluster, and make further study on particular clusters. In addition, cluster analysis usually acts as the preprocessing of other data mining operations. Therefore, cluster analysis has become a very active research topic in data mining. As the development of data mining, a number of clustering methods have been founded, The study of clustering technique from the perspective of statistics, based on the statistical theories, our paper make effort to combine statistical method with the computer algorithm technique, and introduce the existing excellent statistical methods, including factor analysis, correspondence analysis, and functional data analysis, into data mining.

Keywords: Data Mining; Cluster Analysis; Statistical Method

1. Introduction

Since the 90s of the 20th century, with the information technology and the rapid development of database technology, people can Very easy to access and store large amounts of data. The face of large-scale mass data, traditional data analysis work With only some surface treatment, but can not get the inherent relationship between the data and the underlying information, from Fall into the "data rich, knowledge poor" dilemma [1]. To escape this dilemma, people urgently need a species can intelligently and automatically transform the data into useful information and knowledge of techniques and tools, which are on the strong force the urgent needs of data analysis tools make data mining (Data Mining) technology emerged [2].

Data mining in recent years with the database and artificial intelligence developed a new technology that the big amount of raw data to discover the hidden, useful information and knowledge to help policy makers to find the potential between the data Associated factors found to be ignored. Data mining because of its huge business prospect, are now becoming an international data library and information policy-making in the field of cutting-edge research, and caused extensive academic and industry relations note [3]. At present, data mining has been in business management, production control, electronic commerce, market analysis and scientific science and many other fields to explore a wide range of applications [4].

The face of huge amounts of data, the first task is to sort them out, cluster analysis is to classify the raw data as a reasonable way. The so-called clustering is a group of physical or abstract objects, according to the degree of similarity between them, divided into several groups, and makes the same data objects within a group of high similarity, and different groups of data objects are not similar [5][6].

As an important function of data mining, clustering analysis can serve as a stand-alone tool to get data on the distribution of observed characteristics of each class, focus on a specific class to do some further analysis

[+] Corresponding author. *E-mail address*: yujiefeather@sina.com.

[7]. In addition, Cluster analysis can be used as pre-processing steps of other algorithm. Therefore, cluster analysis has become a data mining a very active area of research topic [8].

Against this background, this paper explores statistical perspective on data mining issues in the clustering of deep into the study, the statistical theory, statistical methods and algorithms to the combination of the basic ideas, some of the existing the best statistical methods, such as factor analysis, correspondence analysis, function-based data analysis into the field of data mining, large amount of data it can be used in cluster analysis. Chapter 2 is a data mining and clustering a review. Chapter 3 will be a classic statistical method-Q mode factor analysis into the field of data mining is proposed data mining in the "Q-type factor clustering method. Chapter 4 Benzri correspondence analysis based on the basic ideas, combined with Q-factor analysis on the idea of Chi-square distance of the framework, a new large-scale database clustering method - "Correspondence Analysis Cluster Analysis." Chapter 5 of the paper is summarized and prospects.

2.Cluster Analysis in Data Mining

2.1. The Definition Of Data Mining

Data mining is a large number of incomplete, noisy, fuzzy, random the practical application of the data found in hidden, regularity, people not known in advance, but is potentially useful and ultimately understandable information and knowledge of non-trivial process [9]. "Known in advance" means the information is pre- unanticipated, or novelty. The information unearthed more surprising, the more likely value. "potential the usefulness of the "means of knowledge found in actual effect of future, that information or knowledge of the business discussed or research to be effective, there are practical and achievable [10].

The definition of data mining is closely related to another commonly used term knowledge discovery [11]. Data mining is an interdisciplinary, integrated database, artificial intelligence, machine learning, statistics, etc. Many areas of theory and technology are databases, artificial intelligence, data mining and statistics is a study of three strong large technology pillars.

2.2. The Functions Of Data Mining System

Data mining aims to discover hidden from the database, meaningful knowledge, mainly into the following categories function [12]:

(1). Concept description

Concept description is called as summary description, which aims to concentrate the data, given its comprehensive descriptions, or will compare it with other objects. By summing up the data, you can achieve an overall grasp of the data. Description of the most simplest concept is the use of statistics in the traditional method to calculate the various data items in the database total, mean, variance, etc., or use OL "(On Line Processing, online analytical processing) achieve multi-dimensional query and calculation of data.

(2). Correlation Analysis

Correlation analysis found that large amounts of data items from the set of interesting association or correlation between the contacts. With the large number of continuously collect and store data, and many people in the industry from their database for mining association rules increasingly the more interesting. Records from a large number of business services found interesting correlation can help many business decisions making.

(3). Classification and Prediction

Classification and prediction are two forms of data analysis can be used to extract models describing important data classes or pre-future trends measured data. Classification and Prediction of a wide range of applications, for example, you can create a classification model. On the bank's loan customers to classify, to reduce the risk of the loans; also through the establishment of the classification model the functioning of the factory machines to classify, to predict the occurrence of machine failure.

(4). Cluster Analysis

Category according to maximum similarity and minimize between-cluster similarity principle, makes the same class of objects with high similarity with other classes of objects is very similar. each cluster formation.

Class can be seen as an object class, which it can export rules. Clustering is also easy to observe the contents of the organization into hierarchical structure to organize similar events together.

(5). Outlier Analysis

Database may contain data objects, their general behavior with the data or the model inconsistent. This data objects are outliers. Many data mining algorithms attempt to minimize the impact of outliers, or row. In addition to them, however, in some applications may be an isolated point of a very important message. For example, in fraud detection, isolated points may indicate fraud.

(6). Time Series Analysis

In time series analysis, the data attribute value is changing over time. These data generally equal time intervals to obtain, but can not get equal time intervals. Through the time series map can be time-series data visualization. There are three basic functions in time series analysis: First, dig mode excavation, that is, by analyzing the time series of historical patterns to study the behavioral characteristics of affairs. Second, trend analysis, that is, using historical data for time series forecasting the future value. Third, similarity search, which uses distance measures to ensure given the similarity of different time series.

2.3. Commonly Techniques used Data Mining

Data mining techniques and methods used in the main related disciplines and technologies from the following areas:

(1). Statistical Methods

In data mining often involves a certain degree of statistical process, as data sample and modeling to determine assumptions and error control. Including descriptive statistics, probability theory, regression analysis, time series, including many of the statistical methods, data mining plays an important role.

(2). Decision Tree

Decision tree method is mainly used for data classification. Generally divided into two stages; The tree structure and tree pruning. Firstly, the training data to generate a test function, according to different Classification based on decision tree classification method in comparison with the other, with faster, more easily into simple and easy to understand classification rules, easily converted into database queries advantages, especially in problem areas of high dimension can be very good classification results.

(3). Neural Network

Artificial neural network structure mimic biological god the network is trained to learn through the nonlinear prediction model, in data mining can be used to carry out sub-class, clustering, feature extraction and other operations.

(4). Genetic Algorithm

Genetic algorithm is an optimization technique, which uses students evolution of the concept of property issues a series of search and finally optimized. Implementation of genetic algorithm, the first code for solving problems (called chromosomes), generates the initial population and then calculate the individual fitness, and then chromosome replication, exchange, mutation operation, generate new individuals. Repeat this exercise for, until the individual seeking the best or better. In data mining, data mining tasks tend to express as a search problems, use the powerful search capability of genetic algorithm to find the optimal solution.

(5). Rough Set

Rough set theory is a problem dealing with ambiguity and uncertainty of new mathematical tools, it has a deep mathematical foundation, simple, targeted and computation advantages. Use rough set theory can deal with issues such as data reduction, data correlation was found, meaning the assessment data, the number of according to the approximate analysis [13].

(6). Fuzzy Set

Fuzzy sets is that the uncertainty of data and processing of important ways. Degree of membership of fuzzy set theory to describe the difference with the medium transition is a language with a precise

mathematical fuzziness described method [14]. Fuzzy sets can not only deal with incomplete data, noise or imprecise data, but also in development of data uncertainty models can provide a more agile than traditional methods, smoother performance. Data mining, often used for evidence combination, confidence computing.

3.The Factor Clustering Method

Factor analysis is of the first to study psychology and education issues. Since this study had received good results, causing a lot of statisticians and other experts, attention. Become a classical multivariate statistical analysis, and gradually applied to psychology and education other than school Subjects: economics, sociology, biology, sociology [15].

The traditional view is generally believed that all the calculation process from the perspective, R-type factor analysis and Q-factor scores analysis is exactly the same, but different starting point calculation, R-type factor analysis of correlation coefficients from the variable moment Array starting, Q-type factor analysis from the similarity coefficient matrix of the sample.

3.1. Factor Loading Matrix Using the basic idea of clustering of samples

Clustering method is based on the degree of similarity between each sample. Similarity between samples exists, Because of the different samples are often dominated by a number of common factors, and impact. Q-factor analysis of the purpose is the large number of samples from the impact of the various samples to find more fundamental factors - public factors.

Factor loading matrix of the sample using the basic idea of clustering is by judging the various factors in different samples on the relative size of the load to construct a classification standard: if a certain number of samples on the same factors are relatively large positive load, is illustrated in these samples and the factor also has a strong positive similarity, so these samples can be clustered together; the contrary, if a certain number of samples on the same factor has a greater load bearing, This is illustrated in several samples and the factor also has a strong negative similarity, so these samples can be clustered. Thus, if the extraction of k-factor is, the sample will be divided into zk class.

If the factor loading matrix of different factors in each sample the absolute difference between the load is not very clear, the matrix by the variance of the maximum load rotation, so that each sample only have the absolute value of a public factor larger load, and load factors in the absolute value of the other smaller, so that a clear classification of the sample.

3.2. Data Mining in the Q-factor clustering

We can see from the above analysis, Q-type factor analysis is essentially a similarity coefficient between the sample size to samples for the classification based on clustering method. In fact, as a traditional clustering method, which has been widely applied geology, biology, chemistry, psychology and other fields.

However, as our in-depth study of factor analysis of algorithms will find that to be on a n samples and p variables constitute the nxp matrix of the initial data Q-factor analysis, one must first calculate the nxn covariance matrices, then calculate the characteristic roots of covariance matrices and the corresponding feature vector, and the complex time step algorithm Miscellaneous degree O (mine). In other words, the traditional Q-factor analysis of running time is roughly proportional to mine. Then, when the sample size n large when calculating Q-factor loading matrix to spend a lot of computer resources, therefore, traditional Q-factor analysis does not directly applied to the data mining field. To solve this problem, we try algorithm on Q-factor analysis to improve, so that it can handle huge amount of data clustering problem, and this is known as "Q-type factor clustering method."

4.Correspondence Analysis Clustering

In the last chapter, we analyze the basic ideas with the corresponding improved algorithm for Q-factor model efficiency, has been applied to massive data clustering Q-factor clustering. This chapter we will follow a similar the idea of direct transformation of the correspondence analysis model, to establish another kind of clustering data mining in a correspondence analysis clustering method. Correspondence Analysis

also known as the corresponding analysis, factor analysis in the R-and Q-factor analysis method based developed on, it also has the R-and Q-factor analysis of the characteristics.

In practice, of which, many people use the results of correspondence analysis and cluster samples and variables. General approach is to extract only the first two factors and factor loadings for the past two coordinates, the n points and samples p variable points was drawn on a map Zhang factor loading, and then by observing the sample point and variable point because sub load chart to analyze the relative position of the sample between the sample and variables between the variables and the distance between the pro- close relationship. However, the traditional correspondence analysis to be directly applied to the field of data mining, or at least the following shortcomings.

(1). the traditional two-dimensional correspondence analysis on subjective observation of the load factor to evaluate the relationship between the sample and variables the lack of objective statistics to measure this correlation, it is inevitable subjectivity of the statement obtained.

(2). the sample point and n-p variable points was drawn on a scatter plot Zhang, it is only when the sample variables are not too many goods and when to apply and the field of data mining are frequently faced with thousands of sea amount of data, therefore, this method is almost impossible to use.

(3). to make a scatter plot, usually extracted two factors, the most they can extract three factors about k-dimensional factor loading matrices to two-dimensional or three-dimensional projection space. The projection of the result like due to the real Spaces in the field situation, especially when the original data dimension higher, so the true factor space when the larger dimension k, and data mining are often faced with precisely the high-dimensional data.

4.1. Correspondence Analysis in Data Mining Clustering

To take full advantage of correspondence analysis on the superiority of the algorithm, while avoiding the time factor in extraction caused information loss, and try to construct an objective classification standard cluster samples, we can not for the moment R Type factor loading matrix, which will focus on Q-factor load that came out. Can extract enough because Child, then the corresponding analysis will be the factor loading appropriate transformation, it can truly representative samples and the factor of similarity. At this time, you can use load factor as the classification standard sample clustering. We call this method as "correspondence analysis clustering method."

4.2. Correspondence Analysis Clustering Method In The Mobile Communications Market Segments

Market segmentation is the telecom carrier market marketing efforts. First of all, by user characteristics, consumption habits the breakdown of such dimensions, the general user market segmentation will become a significant feature of a number of market segments, so that the same fine between the individual sub-markets to minimize the inherent differences make the difference between the different market segments to the most great, then you can combine features of different market segments to develop specific strategies to effectively develop and meet specific market and consumer demand. So far, this article has introduced the two clustering methods: Q-type factor clustering and correspondence analysis clustering method. We can see that the two methods in terms of ideology or in the algorithm, there are a lot of similarities. For instance, both methods are time complexity of the sample size of the linear stage, two kinds of methods to Q-Factor load matrix for clustering based on factor scores are used to explain features of the category, and so on. Of course, both methods have some obvious differences:

(1). although the two methods in Q-factor loading matrix as a cluster basis, but the measure of both space is not like Q-factor clustering method is based on European metric space, while the corresponding analysis of clustering.

(2). Secondly, although both methods are low by solving a matrix of characteristic roots and characteristic vectors and thus inter- ground by Q-factor loading matrix of ways to improve the efficiency.

(3). Finally, Q-type factor analysis clustering method can only be applied to quantitative data analysis, and the corresponding analysis not only can deal with quantitative data, qualitative data can deal with.

4.3. Comparison and Analysis of Algorithms

With the existing clustering methods for comparison, we combine the standard from the cluster, Class identification and algorithm framework proposed in paper three aspects of the two clustering methods to make a brief summary.

(1). Clustering criteria

Q-factor analysis is essentially a kind of similarity coefficient between the size of the sample classification based on clustering method therefore, both the Q-factor clustering or clustering correspondence analysis, are based on similarity coefficient clustering standard, but the former is based on the calculation of similarity coefficient European metric space, which is based on the chi-square distance air similarity coefficient between the calculated. Similarity coefficient can be regarded as a special kind of distance metrics. Therefore, clustering the standard point of view, these two methods presented in this paper belong to a distance of standard Quasi-clustering method. As mentioned above, in order to distance the standard clustering method as the clustering of abnormal points are usually higher than Sensitive, therefore, how to effectively exclude the impact of outliers is the two methods require further study.

(2). Class identity

In the Q-factor clustering and correspondence analysis clustering method, all samples will be divided into Zk categories, gathered in the k-factor axis in the direction around the positive and negative, so the two clustering methods are a common factor with k. Positive and negative direction as a class identity, which is obviously a single representative point for the type of clustering method identified, and here is representative of the original data points do not exist. Single representative point method is usually only recognizes convex or spherical class, this defect also exists in the proposed two clustering methods.

(3). Algorithm Framework

Existing clustering methods can be classified according to their algorithm framework optimization, search, gather and split the four categories. This article proposed Q-factor clustering and correspondence analysis clustering method, speaking from the algorithms, they all inherit the factor Analysis. The factor analysis itself is an optimized method, which is reflected in its two main processes: factor loading matrix of the solution process is essentially a process of extreme demand conditions; factor rotation processes makes the variance is a maximum load factor optimization process.

5.Conclusion

Data mining in recent years with the database and artificial intelligence developed a new technology, its aim the large amount of data from the excavated useful knowledge, to achieve the effective utilization of data resources. As one important function of data mining, clustering analysis either as a separate tool to discover data sources distribution of information, as well as other data mining algorithms as a preprocessing step, the cluster analysis has been into the field of data mining is an important research topic.

From the statistical perspective on the problem of clustering data mining in-depth study to statistical theory Based on statistical methods and algorithms to integrate the basic idea, put forward some new clustering method, and the clustering method was successfully applied to the social, economic, management and other fields.

6. Acknowledgment

This work is supported by GuangXi Economic Management College Foundation the Project Number is: 08CF0101005

7. References

[1] Esehrieh S. Jingwei Ke, Hall L. O. etc. Fast accurate fuzzy clustering through data reduction [J].IEEE Transactions on Fuzzy systems，2003，11(2):262-270.

[2] Zahid N，Abouelala O，Limouri M，etc. Fuzzy clustering based on K-nearest-neighbours rule[J].Fuzzy Sets and Systems，2001，120(2).

[3] Strehl A，Ghosh J. Relationship-based clustering and visualization for high-dimensional data mining[J].INFORMS J COMPUT，2003，15(2):208-230.

[4] Milenova B.L.， Campos M.M.O-Cluster: scalable clustering of large high dimensional data sets[C].IEEE International Conference on Data Mining，2002，290-297.

[5] Daniel B.A.， Ping Chen Using Self-Similarity to Cluster Large Data Sets[J].Data Mining and Knowledge Discovery，2003，7(2):123-152.

[6] Wei Chi-Ping， Lee Yen-Hsien， Hsu Che-Ming. Empirical comparison of fast Partitioning-based clustering algorithms for large data sets[J].Expert Systems with Applications，2003，24(4):351-363.

[7] Maharaj E. A. Cluster of Time Series[J].Journal of Classifieation，2000，17(2):297-314.

[8] Guedalia I.D.， London M.， Werman M. An on-line agglomerative clustering method for non-stationary data[J].Neural Computation，1999，11(2).

[9] Jung-Hua Wang，Jen-Da Rau，Wen-Jeng Liu. Two-stage clustering via neural networks[J]. IEEE Transactions on Neural Networks，2003，14(3):606-615.

[10] Veenman C.J.，Reinders M.J.T.，Backer E. A maximum variance cluster algorithm，[J].IEEE Transactions on Pattern Analysis and Machine Intelligence，2002，24(9):1273-1280.

[11] Cattell Raymond B. The Three Basic Factor Analytic Research Designs-Their Interrelations and Derivatives[J].Psychological Bulletin，1952，49:499-520.

[12] Stephenson，william. Some Observations on Q Technique[J].Psychological Bulletin，1952，49:483-498.

[13] Riehard A.，Johoson，Dean W.，Wichern. Applied Multivariate Statistical Analysis(5 th Ed) 2003.

[14] Guttman L. The quantification of a class of attributes: A theory and Method of scale construction[C].The Committee on Social Adjustment(ed.)，The Prediction of Personal Adjustment. New York : Social Science Research Council，1941.

[15] Bnzecri, J. P. Statistical analysis as a tool to make Patterns emerge from data[C].In S. Watanabe(ed.)，Methodologies of Pattern Recognition. New York: Academic Press，1969:35-74.