Journal of Theoretical and Applied Information Technology
15th December 2012. Vol. 46 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ANALYSIS ON ALGORITHM AND APPLICATION OF CLUSTER IN DATA MINING

YUHUA FENG

Information Engineering School of Nanchang University, Nanchang 330031, Jiang xi, China

ABSTRACT

Cluster analysis in data mining is an important research field; it has its own unique position in a large number of data analysis and processing. The research about Clustering makes a spurt development after more than 20 years, and then produced a variety of types and application of the clustering algorithm. Data Mining is one of the pop researches in information industry last few years. This paper analyses some typical methods of the cluster analysis and represent the application of the cluster analysis in Data Mining.

Keywords: *Data Mining, Cluster Analysis, Cluster Algorithm, K-Means*

1. INTRODUCTION

Because of the rapid development of information technology in recent years, the large amount of data to be stored in the database, how to analysis the effective data and tap the potential of information has become the focus of research. The technology of mining knowledge from large amounts of data called data mining (Data Mining, DM). the statistics is fundamental for Data mining, clustering analysis as one of three methods for multivariate data analysis is the core technique of data mining, so clustering analysis in data mining is an important research field, it can be used as an independent tool to access the distribution of data in database, but also can be used as other data mining analysis algorithm in a preprocessing step. Now Clustering analysis in data mining has become a very active research topic.

2. ANALYSIS OF CLUSTER ANALYSIS ALGORITHM

Since 80's the clustering algorithm began to process data, the research of clustering algorithms has not stopped. Facing the different application requirements, the researchers suggest many kind of clustering algorithm, so it can be seen in clustering algorithm has many applications. The clustering algorithm can be defined by a simple to describe "through clustering operation, data object is divided into subsets in order to as far as possible similarity in the same subset and as far as dissimilar in the different subset."

2.1 Concepts of Clustering

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. We can show this with a simple graphical example:

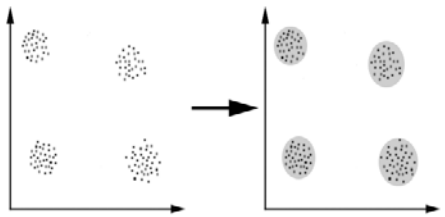

Figure 1: A Simple Graphical Example

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Cluster analysis originates "like attracts like "the simple idea, so the clustering or classification is according to the data characteristic. So a high quality clustering algorithm must satisfy the two

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

conditions: the similarity of data or the object in same subset is the strongest; the similarity of data or object in different subset is the weak. The quality of clustering usually depends on similarity measuring method and the realization of the way used by the clustering algorithm, but also depends on the algorithm it can find all or part of the hidden pattern.

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection).

2.2 Analyses of Clustering Algorithms

Now summarizing the existing clustering methods, we have boiled it down to the following categories: (1) Partitioning Methods, a given database that contains n data object or tuple generates the number for the K cluster partition based on clustering algorithm. The standard of Division (or known similarity function) usually called Euclidean distance, for the categorical data attributes you can use the Jaccard coefficient. All K-means and Fuzzy C-means is the most famous two in this method, these clustering methods are very applicable for finding globular clusters mall in medium database. But in order to clustering for large-scale data set, and clustering complex shape, the method needs to be further expanded.

(2)Hierarchical Methods, the method is decompose a given data object set for many levels, so it will build a clustering tree. According to the hierarchical decomposition is based on bottom-up or top-down principle, it can be further divided into condensed and division. The former to each individual as a separate class and it process with the data similarity, then a sufficiently large data gradually merged into larger categories; the latter conversely, the entire set as a category, and then gradually divided into small different types. In order to make up for no traceable deficiency of decomposition or polymerization, hierarchical clustering method often combine some other methods, such as circular positioning. The typical hierarchical clustering methods such as BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)

(3) Density-based Methods, it see cluster as a high density area what is splitting a space by low density region. The basic idea is following: as long as the adjacent region of the dot density (number of data points) beyond a certain threshold, will continue to clustering, till the field must contain at least a certain number of points. The algorithm is based on the sample of neighborhood conditions, the whole sample space is split by low density interval, and it does not need to know the cluster number in advance, it complete the clustering task through one scan. The difference of Density-based Methods and other methods of a fundamental is that it is based on the density not on the variety of distance, so it can overcome the situation that the algorithm based the distance can only be found in "dough " cluster. There are some typical algorithms such as OPTICS (Ordering Points to Identify the Clustering Structure)

(4) grid-based method, this method first is divided data space into a grid unit, and then mapped the data set of samples into the grid cell; each cell density will be calculated. According to given density threshold to judge whether each grid unit is high density unit, so several adjacent dense grid unit formed clusters. The main advantage of this approach is fast processing speed, which independent of the number of data objects, but is with each dimension unit number of the quantization space. At present the common grid clustering algorithm includes Wave Cluster 。

(5) model-based method, the method assumes a model for every cluster, then looks for the model data sets it can well meet the model. Such as through the construction of density function that reflecting the spatial distribution to achieve clustering, its theory is that the data is generated according to the underlying probability distribution , this clustering method can optimize the given data and certain mathematical model more adaptability, such as RSDE algorithm.

(6) Spectral-based Clustering, The method convert the data set into an undirected connected graph, the sample of data set is Vertex data of diagram, the edge weights can reflect the degree of similarity between the graph vertices. Then the clustering problem transfers into the graph partition problem. The best division effect is the weighted summation of maximum spanning between internal vertex of a sub graph, and minimize the weight of between the vertices of a graph, such as GRC (Graph-based Relaxed Clustering)

Journal of Theoretical and Applied Information Technology
15th December 2012. Vol. 46 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

So we can see there are many clustering algorithms, every method has its characteristics, and every kind of clustering algorithms can be seen its own useful in many application.

2.3 One example of Clustering Algorithms

Now, there is an example of clustering algorithms for K-means. For the following table 1, we will divide the data into 2 clusters using K-means, and assume that the initial cluster centers selected for P7 (4, 5), P10 (5, 5).

Table 1: Sample data

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
x	3	3	7	4	3	8	4	4	7	5
y	4	6	3	7	8	5	5	1	4	5

(1) The following distance calculation of a sample of 10 to 2 of this cluster center, and assigned 10 samples to its closest cluster.

(2) The first iteration of the results are as follows：

Belonging to the cluster C1 samples：{P7，P1，P2，P4，P5，P8}.

Belonging to the cluster C2 samples：{P10，P3，P6，P9}.

Then to compute new cluster center：the center of C1 is (3.5，5.167)，and the center of C2 is (6.75，4.25)

(3) It continues to calculate the distance of the 10 samples to new cluster center, and then assign them to new cluster. The second iteration of the results is as follows:

Belonging to the cluster C1 samples {P1，P2，P4，P5，P7，P10},

Belonging to the cluster C2 samples {P3，P6，P8，P9}.

Now we get the new cluster center: the center of C1 is (3.67，5.83)，and the center of C2 is (6.5，3.25)

(4)Continue to calculate 10 samples to the new cluster center distance, reassigned to the new cluster, cluster center that will not change the algorithm terminates.

K-means algorithm described in easy, simple, fast, but there is insufficient:

- The number of clusters is difficult to determine;
- The clustering results is sensitive to the selection of initial values;
- The algorithm uses climbing type technology to find the optimal solution, it is easy to fall into local optimal value;
- The algorithm is sensitive to noise and abnormal data;
- It cannot be used for non convex cluster, or with a variety of different size cluster.

3. THE APPLICATION OF CLUSTER ANALYSIS IN DATA MINING

The application of cluster analysis in data mining has two main aspects: first, clustering analysis can be used as a preprocessing step for the other algorithms such as features and classification algorithm, and also can be used for further correlation analysis. Second, it can be used as a stand-alone tool in order to get the data distribution, to observe each cluster features, then focus on a specific cluster for some further analysis. Cluster analysis can be available in market segmentation, target customer orientation, performance assessment, biological species etc.

3.1 The Application of Text Clustering

Text clustering is an important application for clustering algorithm; it emerged from text retrieval, and had important application in establishing meaning network, information retrieval, knowledge management system. Because of large quantities data and various types' characteristics of text clustering, the text clustering research is easy to expand to other application fields.

English text is composed by words, and the Chinese text is composed by terms, but other language may have new semantic unit. So far, for the natural language text, we still can't use strict syntax and semantic rules to analyze text semantic that classified the text comparative intelligently.

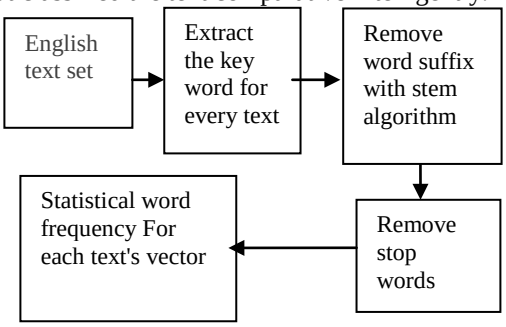

Figure 2: Pretreatment Processes For Text Clustering Algorithm

Due to the nature of the English text itself, English words are separated by Spaces, and basically every word can express the meaning independently. Therefore, the key word for extraction is space, all kinds of punctuation is the limits of the document words.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

For the Chinese, a simple method is to extract characters, but it makes the document originally expression of semantic caused great destruction, so it processes Chinese document keywords that it always use segmentation technology for the Chinese word, which express a complete semantic with some words in this problem.

Document used vector model to represent document which is applied widely in the information retrieval field. The p_i is expressed as $p_i = (w_{i1}, w_{i2}... w_{ij}... w_{im})$, and m as the total of key word, w_{ij} is the weight of the pi vector in Numbers for j keywords for document p_i.

$w_{ij} = freq_{ij} \times \log(N/dfreq_j)$

Among them, the $freq_{ij}$ is the occurrence number of the key word j in document i, $dfreq_j$ is the document number includes key word j in the document, N is the size for document set.

3.2 Some Relative Applications

The cluster analysis has been applied to many occasions. For example, in commercial, cluster analysis was used to find the different customer groups, and summarize different customer group characteristics through the buying habits; in biotechnology, cluster analysis was used to categorized animal and plant populations according to population and to obtain the latent structure of knowledge; in geography, clustering can help biologists to determinate the relationship of the different species and different geographical climate; in the banking sector, by using cluster analysis to bank customers to refine a user group; in the insurance industry, according to the type of residence, around the business district, the geographical location, cluster analysis can be used to complete a automatic grouping of regional real estate, to reduce the manpower cost and insurance company industry risk; in the Internet, cluster analysis was used for document classification and information retrieval etc.

4. SUMMARY

Scholars and industry users attach more and more importance on Clustering algorithm as one of the most important technologies for data mining.

As a result of data collection ability and hasty study in the industry business analysis system, clustering analysis model have yet to be further research and perfection in each group of the division and Variable selection and so on ,We can see there is a lot of work to do for f the cluster analysis, which the application of cluster analysis bring convenience to the people. For example, through market research, it can strengthen customer investigation and research work to get more detailed and comprehensive customer consumption psychology, the features of attitude of some aspects of the variables. The application of cluster analysis is more and more urgent; the requirements are also getting higher and higher. With the development of modern technology, in the near future, cluster areas will achieve a critical breakthrough.

REFERENCES:

[1] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms", *Computer Journal*, Vol. 26, No. 4, 1983, 354-359.

[2] J. Yang, W. Wang, P. S. Yu, and J. Han, "Mining long sequential patterns in a noisy environment", *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, ACM Press, 2002, pp. 406–417.

[3] J. S. Park, M. S. Chen, P. S. Yu, "An effective hash based algorithm for mining association rules", *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, California, 1995, pp.175–186.

[4] G. Blanchard, G. Lugosi and N. Vayatis, "On the rates of convergence of regularized boosting classifiers", *J. Machine Learning Res.*, Vol. 4 2003, pp. 861–894.

[5] J. Hartigan, M. Wong, "Algorithm AS136: A k-means clustering algorithm", *Applied Statistics*, Vol. 28, 1979, pp. 100-108

[6] B. Liu, Y. Xia, and P. S. Yu, "Clustering through decision tree construction", *Proceedings of SIGMOD 2000*.

[7] J. Mao, J. A. K. Jain, A Self-organizing network for hyperellipsoidal clustering(HEC), *IEEE Transactions on Neural Networks*, Vol. 7, No. 1, 1996, pp. 16-29.

[8] I. Sarafis, A.M.S .Zalzala, P.W.Trinder, "A genetic rule-based dataclustering toolkit", *Proceedings of 2002 Congress on Evolutionary Computation*, 2002, pp. 1238-1243.