Clustering Algorithms In Data Mining

Xiaosong Chen [1, a]

[1]Deparment of Computer Science, University of Vermont, Burlington, VT 05401, USA

[a]xchen20@uvm.com

Keywords: Data mining, Algorithm, Clustering.

Abstract. Data mining is a hot research direction in information industry recently, and clustering analysis is the core technology of data mining. Based on the concept of data mining and clustering, this paper summarizes and compares the research status and progress of the five traditional clustering algorithms, and sorts out some new clustering algorithms. Nowadays, clustering algorithms are mainly focused on fuzzy clustering, spectral clustering, quantum clustering and so on. This study is a good summary of the clustering algorithm, and it has positive significance for the development of clustering.

1. Introduction

With the development of database technology and the rapid popularization of the Internet, the amount of data people facing has increased dramatically, and commercial, business, scientific research institutions or government departments have accumulated a large number of data stored in different forms. However, at the same time, we are lack of understanding and application of fully contained in the data information and knowledge, relying on the traditional database for the data query, retrieval and other analysis methods can not help users to extract useful information with the conclusion from the data, nor can meet the requirements of data processing and analysis. Under this circumstance, the Knowledge Discovery in Database (KDD) and Data Mining (DM) technology came into being and showed strong vitality [1, 2]. Data mining is an interdisciplinary branch of computer science that uses the artificial intelligence, machine learning, statistics, and database cross-over methods to calculate the discovery patterns in relatively large data sets [3]. The practical work of data mining is the automatic or semi-automatic analysis of large-scale data to extract the valuable information of the past unknown [4]. The key of data mining is trying to discover the valuables from database, and to achieve it, there are many related techniques used. For instance, the data is grouped by cluster analysis, and the abnormal records of data are detected by anomaly detection, and the relation between data is mined by association rule mining.

Since data mining is obtained from the basis of database, machine learning, database development and mathematical statistics technology, the clustering technique, as a major technical and functional technique of data mining, has become a very active research field. What are clustering techniques? The definition work here for is dividing data into some many smaller groups, contain in similarity (dissimilar to items in other groups), from database. This is also familiar called unsupervised algorithm, using actively in statistics, machine learning, data mining, and medical biology research field. In the field of statistical analysis and pattern recognition, clustering has been widely studied for many years, and a large number of theories and methods have been put forward, and fruitful results have been obtained. At present, clustering has been widely used in various fields of engineering and science, such as psychology, biology, medicine, etc.

2. Classification and comparison of classical clustering algorithms

Clustering analysis techniques can be divided into five categories, which are partition based clustering, hierarchical clustering, density based clustering, mesh based clustering, and model-based

clustering. There are many specific algorithms in these five categories, and the related clustering algorithms will be introduced later.

2.1 Partition based method

The most classic and popular data partitioning relocation clustering algorithm are K-means and K-medoids methods, which represent each cluster C_j by the mean c_j of its points, the so-called centroid. The K-means algorithm accepts a parameter K to determine the number of clusters in the result. At the beginning of the algorithm, the K data objects are randomly selected in the data set to be used as the initial centers of the K clusters, and the remaining data objects are selected from the distance of the cluster to each cluster heart, and assigned to them. Then, the average values of all the data objects in each cluster are recalculated, and the result is taken as the new cluster heart. The process is repeated gradually until the objective function converges. Usually, the mean square function is used as the objective function. The formula is as follows:

$$J = \sum_{i=1}^{k} \sum_{D \in C_i} \left| D - m_i \right|^2 \tag{1}$$

Where J is the sum of the mean square of all the data in the data set and the corresponding clustering center, D is the data object, and m_i is the average of the cluster C_i.The advantage of this algorithm is simple and fast enough to process data, whereas it limitation for seeking mean of the point inside a cluster is not easily get.

For K-medoids, is represented by one of points in the cluster. Once the medoid is selected, the objective funiction is defined as average distance between point and medoid, this is the difference between K-mean method. Its disadvantage is less sensitive to the dirty data and unusual data, so that it work better in small size data circumstance.

K-means and K-medoids algorithm are the two classic algorithms in the division algorithm. Many other partitioning algorithms have been improved from the evolution of these two algorithms

2.2 Hierarchical approach

Hierarchical clustering is literally referring to its name, a tree structure of clusters and for every cluster have its own parent (except the top priority cluster), child and sibling clusters. This approach allows visiting data clusters in different levels. Generally using method include the bottom-up, start with single cluster and recursively merges more clusters, and top-down, start with total clusters and recursively splits into several clusters. The final goal is to achieve the requirement of specific number k of clusters. This method is flexible and doing well in handling any forms of similarity. But lack of a terminal criteria for vague data.

The basic hierarchical clustering methods are aggregation and splitting methods proposed by Kaufman and Rousseeuw. The drawback is that merging points or split points are more difficult to select and less scalable. CURE (Clustering Using Representatives) method proposed by Guha does not use a single center or object to represent a class, but to choose a fixed number of representative points in a data space to represent a class. It can identify different classes with complex shapes and sizes, but also can filter isolated point [5]. Based on the CURE method, Guha proposed the ROCK method for classifying data [6].

The advantage of the hierarchical clustering algorithm is that the algorithm can get the multi-level clustering structure with different granularity, but there are the following shortcomings: (1) it is often difficult to merge or split the point selection. If the merger or splitting is not good, it may lead to a decline in the quality of clustering. (2) since the merger or split need to check and estimate a large number of samples or clusters, algorithms are not ideal for scalability. (3) the algorithm time complexity is high, which is $O(n^2 \lg n) \sim O(n^3)$

2.3 Density-based Method

The main idea of density-based clustering method is: set a sample in the space as the center, per unit volume of sample number is called the point density, intuitively, the cluster sample density is high, density of inter cluster sample is relatively low. According to the difference of spatial density, the algorithm regards cluster as a high density sample region separated by low density region in

sample space. The clustering continues as long as the density of the adjacent region (the number of objects or data points) exceeds a threshold. The difference between the density algorithms is mainly about how to define the high density region and the low density region.

Before implementing the Density-Based Partitioning for a certain dataset, concept of density and connectivity and boundary are prereqs. Central idea of this algorithm is that once any one of the points is greater than prescribed threshold, then Partitioning it into the similar density cluster. Density-Based algorithm is capable of exposing the covered clusters of arbitrary shapes. However, the limitation of it is only effective work in low-dimensional data, since underlying attribute required space.

The most typical density based algorithm is DBSCAN [7]. The main goal of the density-based clustering algorithm is to find high density regions separated by low density regions. Different from the distance-based clustering algorithm, the results of clustering algorithm based on distance is globular clusters, and the density based clustering algorithm can discover clusters of arbitrary shape, which plays an important role to the noisy point data.

2.4 The Grid-Based Method

For Grid-Based Method, firstly divide the sets of data into finite segments (a Cartesian product of individual sub-ranges) of unit structures. This method is good at the processing in really fast speed. This is not relating to the number of targeted records in database, but how many units it contains. Data partitioning is induced by point's membership in cell resulted from space portioning, the accumulation of grid-based data makes clusters independent of ordering. The disadvantage of this algorithm is that the user is required to give a density threshold λ, and the threshold setting has a great effect on the clustering result. If the threshold λ is too large, some clusters may be lost. If the threshold setting is too low, it may merge adjacent clusters, and it is not easy to set different thresholds if the density of the data clusters is different or the noise density is different. Therefore, the grid-based clustering analysis algorithm is faced with the balance between computational complexity and computational precision and the number of cells. If the number of cell meshes is too small, the accuracy of clustering results will be weakened. If the number of cell meshes is too large, the complexity of the algorithm will be improved. Common representative grid-based clustering algorithms are Wave Cluster [8], CLIQUE [9] and so on.

2.5 Model-based Method

The model-based approach assumes a model for each class, looking for the best fit of the data for a given model. Such a method is often based on the assumption that the data is generated based on a potential probability distribution. There are two main methods of model-based methods: statistical methods and neural network methods.

Concept clustering is a kind of clustering method in statistical method, which produces a classification model of data objects, and gives a description of each classification model, that is, each class represents a concept. It uses probabilistic measures to determine concepts and describes the concepts derived. The typical concept clustering algorithm is COBWEB algorithm [10], which is an incremental concept clustering algorithm for categorical attribute data. The algorithm creates hierarchical clustering in the form of a classification tree, and uses a heuristic estimation - classification utility to guide tree construction. COBWEB assumes that the probability distribution on each attribute is independent, and in practice this assumption is not always valid. In addition, the probability distribution representation of clustering cost the updating and storage clustering a great deal, especially when the attributes have a large number of values.

The neural network method describes each cluster as a prototype. According to some distance metrics, the new object is assigned to the class represented by the prototype that is most similar to it. The attributes assigned to an object of a class can be predicted based on the prototype attribute. The representative approach is the Competitive Learning algorithm and the Self-Organizing Feature Map algorithm.

3. Development trend of cluster analysis

Currently, clustering techniques are ubiquitous well using in scientific, industrial and business area. However, because each of the traditional clustering methods is flawed, coupled with the complexity of the actual problem and the diversity of data, so that one method can only solve one certain type of problem. In recent years, with the continuous development of traditional methods in artificial intelligence, machine learning, pattern recognition and data mining, as well as the emergence of new methods and new technologies, the clustering analysis method in data mining has been developed by leaps and bounds. On the whole, it mainly focuses on fuzzy clustering, spectral clustering, quantum clustering and other aspects of research.

3.1 Uncertainty Clustering

In practice, most of the objects do not have strict attributes, their genre and morphology exist intermediary, which are suitable for soft division. The uncertainty clustering method based on objective function combines the clustering into a constrained nonlinear programming problem, and obtains the uncertainty partition and clustering of the data set by optimizing the solution. The method is simple in design, wide in solving the problem, and can be transformed into optimization problem and solved by classical mathematical nonlinear programming theory and is easy to be realized on the computer. Therefore, with the application and development of computer, uncertainty clustering algorithm based on objective function has become a new research hotspot.

3.2 Spectral Clustering

In order to cluster in any shape of the sample space and converge to the global optimal solution, the scholars began to study a new class of clustering algorithm called spectral clustering algorithm. BUHMANN JM [11] first proposed a spectral clustering algorithm, which is based on the theory of spectral graphs and clusters with the eigenvectors of similarity matrices of data, making the algorithm independent from the dimension of the data points but related to the number of data points, which collectively referred to as spectral clustering method. The spectral clustering algorithm is a method based on the similarity between two points, which makes the method applicable to non-measure space. Compared with other methods, this method is not only simple, easy to be implemented, easy to fall into the local optimal solution, but also has the ability to identify non-convex distribution clustering ability, which is very suitable for many practical application problems.

3.3 Quantum Clustering

Quantum mechanics is a science that studies the distribution of particles in the energy field. Clustering is the distribution of the samples in the scale space. It can be seen that the distribution of the particles in the space studied by quantum mechanics is the same as that of the clustering studies in the scale space. As for quantum clustering with known distribution of the sample, the wave function describing the particle distribution is known for quantum mechanics. The clustering process is: when the wave function is known, the Schrodinger equation is used to solve the potential energy function, and this potential function will ultimately determine the particle distribution. Many examples show that the algorithm is satisfied with the clustering problem that the traditional clustering algorithm cannot solve.

4. Conclusions

Cluster analysis is an important research field in data mining. It is a kind of processing method of dividing or sorting data. At present, the research on clustering analysis algorithm has been involved in databases, data mining, statistics, uncertainty mathematics and other disciplines and has made great achievements in development. The clustering analysis algorithm can be divided into many kinds according to different classification methods. In this paper, the clustering analysis algorithm is divided into five categories: partition based method, density-based method, hierarchical method, grid-based method and model-based approaches. This paper analyzes and compares the advantages and disadvantages of these different algorithms, and gives the development trend of clustering

algorithm in data mining. The clustering results of clustering algorithm have some unpredictability, in practical application, we should choose the appropriate clustering algorithm according to the data type, in order to obtain the best clustering effect.

References

[1] Piatetsky-Shapiro, G. (1996). Advances in knowledge discovery and data mining (Vol. 21). U. M. Fayyad, P. Smyth, & R. Uthurusamy (Eds.). Menlo Park: AAAI press.

[2] Azevedo, A. (2015). Data Mining and Knowledge Discovery in Databases. InEncyclopedia of Information Science and Technology, Third Edition (pp. 1713-1722). IGI Global.

[3] Clifton, C. (2010). Encyclopedia britannica: definition of data mining. Retrieved on, 9(12), 2010.

[4] John Lu, Z. Q. (2010). The elements of statistical learning: data mining, inference, and prediction. Journal of the Royal Statistical Society: Series A (Statistics in Society), 173(3), 693-694.

[5] Guha, S., Rastogi, R., & Shim, K. (1998, June). CURE: an efficient clustering algorithm for large databases. In ACM Sigmod Record (Vol. 27, No. 2, pp. 73-84). ACM.

[6] Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. Information systems, 25(5), 345-366.

[7] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd(Vol. 96, No. 34, pp. 226-231).

[8] Sheikholeslami, G., Chatterjee, S., & Zhang, A. (2000). WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. The VLDB Journal—The International Journal on Very Large Data Bases, 8(3-4), 289-304.

[9] Domeniconi, C., Papadopoulos, D., Gunopulos, D., & Ma, S. (2004, April). Subspace clustering of high dimensional data. In Proceedings of the 2004 SIAM international conference on data mining (pp. 517-521). Society for Industrial and Applied Mathematics.

[10] Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. Machine learning, 2(2), 139-172.

[11] Arbib, M. A. (Ed.). (2003). The handbook of brain theory and neural networks. MIT press.