Research on the application of clustering algorithm based on minimum spanning tree

Chen Ye[*]

College of Science and Technology, Ningbo University, Ningbo City, Zhejiang Province, China, 315212

Received 28 October 2014, www.cmnt.lv

Abstract

Cluster analysis is one of the most important technologies in data mining. Minimum spanning tree (MST) is an advanced algorithm in cluster analysis. Studying MST has important practical significances. Firstly, this paper analysed partitioning, hierarchical, density and grid clustering algorithms based on MST thoroughly. Secondly, implementation principles and shortcomings of these four algorithms were discussed. Finally, practical applications of clustering algorithm based on MST were introduced, aiming to solve some practical problems.

Keywords: data mining, clustering algorithm, MST model

1 Introduction

Nowadays, mature database technology has been developed and data application has been promoted to a new high. People are facing with big data every day. To utilize such big data effectively, researchers developed knowledge discovery in database (KDD) and improve it continuously. Various data mining technologies for different algorithms have been developed [1].

Cluster analysis is an important technology of data mining. However, clustering technology is still developing. Research on cluster analysis not only has theoretical significance, but also can widen its practical applications. Figure 1 shows the whole process of cluster analysis.

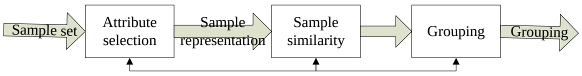

FIGURE 1 Process of cluster analysis

As things of one kind come together, researchers divide data objects into different groups according to their attributes. This is known as cluster analysis (Figure 2)

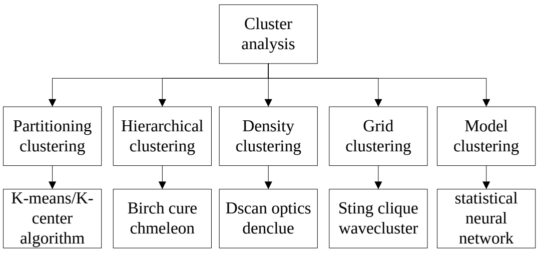

FIGURE 2 Cluster analysis

To overcome shortcomings of classical clustering analysis algorithms, a clustering algorithm based on MST was developed. It is an advanced cluster analysis and can eliminate blindness of clustering analysis significantly. With the continuous development and improvement of MST, four clustering algorithms based on MST have been developed: partitioning clustering algorithm based on MST, hierarchical clustering algorithm based on MST, density clustering algorithm based on MST and grid clustering algorithm based on MST [2] (Figure 3).

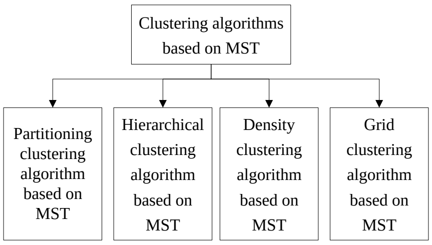
Clustering algorithms based on MST
Partitioning clustering algorithm based on MST
Hierarchical clustering algorithm based on MST
Density clustering algorithm based on MST
Grid clustering algorithm based on MST

FIGURE 3 Classification of clustering algorithms based on MST

2 Classification of clustering algorithms based on MST

2.1 PARTITIONING CLUSTERING ALGORITHN BASED ON MST

Principle: in the partitioning clustering algorithm based on MST, features of dataset D are analysed and its complete undirected graph with weight (G) is constructed firstly. Secondly, researchers will divide the MST (T) into k subtrees according to the characteristics of G. So an initial cluster valued k is gained. When dividing the T, there's a common practice: longest k-1 sides in T will be found and deleted, so that the distance sum of sides in all subtrees will be the smallest and k initial clusters are acquired. This is based on the principle that different points can be grouped into different clusters according to different side lengths. In D, distance between sides of different clusters is longer than distance between points within the same cluster.

[*] ***Corresponding author*** e-mail: yechen@nbu.edu.cn

Therefore, we will get a globally optimal solution. Finally, *k* clusters can be generated by k-means clustering and adjusting affiliation of data objects in different clusters.

Steps of partitioning clustering algorithm based on MST are:

Step 1: Users construct the complete undirected graph with weight (*G*) of the clustering dataset (*D*) according to its characteristics.

Step 2: Users build the MST (*T*) of *G* by using appropriate algorithm.

Step 3: Users set parameters and divide *T* into *k* subtrees according to a specific criterion. *k* initial clusters and the initial cluster centre are generated.

Step 4: Users adjust the cluster centre continuously until get the desired one.

To sum up, partitioning clustering algorithm based on MST has significantly higher accuracy than other classical ones. This is contributed by its scientific and practical generation of initial cluster. Instead of random choose, it generates *k* initial clusters according to the principle of MST. Although the partitioning clustering algorithm based on MST has higher scientific and practical value than traditional ones, it still has some shortcomings for its intrinsic properties:

1) Low generation efficiency of *T*. The time complexity of *T* generation is the function of n^2: $0(n^2)$.

2) Low flexibility. When dividing the *T*, we have to delete the longest *k* sides. This is impractical for many datasets with different geometric distributions. In Step 3, the actual geometric distribution of dataset shall maintain close to the initial cluster in order to get good clustering effect. However, we have to set some parameters and use different division rules. Therefore, improvement is needed to make improved MST satisfy actual cluster distribution [3].

2.2 HIERARCHICAL CLUTERING ALGORITHM BASED ON MST

Analysis: Users build MSTs of the clustering dataset. Later, they cluster data objects using different hierarchical processes according to some characteristics of these MSTs.

Principle: Hierarchical clustering algorithm is an important algorithm. It is composed of agglomeration and splitting. Hierarchical clustering that decomposes from bottom to up is called as agglomeration. In agglomeration, each object is viewed as an independent class. Then, features of these objects will be analysed and similar classes will be combined. It ends until classes are combined into one class. Hierarchical clustering that decomposes from upper to bottom is known as splitting. It concentrates all data objects together as a class. Later, these data objects will be iterated continuously to split the class into smaller classes. It ends until each small class contains only one data object. Both agglomeration and splitting enable users to set different end conditions according to desired classes. Distance between classes is an important reference index to agglomeration and splitting. Common measurements of distance between classes are listed:

Minimum distance:

$$d_{\min}\left(c_i - c_j\right) = \min_{p\in c_i, p^{'}\in c_j}\left|p - p^{'}\right|. \tag{1}$$

Maximum distance:

$$d_{\max}\left(c_i - c_j\right) = \max_{p\in c_i, p^{'}\in c_j}\left|p - p^{'}\right|. \tag{2}$$

Mean distance:

$$d_{mean}\left(c_i - c_j\right) = \min_{p\in c_i, p^{'}\in c_j}\left|m_i - m_j\right|. \tag{3}$$

Average distance:

$$d_{avg}\left(C_i, C_j\right) = \frac{1}{n_i n_j}\sum_{p\in C_i}\sum_{p^{'}\in C_j}\left|p - p^{'}\right|. \tag{4}$$

In hierarchical clustering algorithm based on MST, users build MSTs according to different features of the clustering dataset, which then will be put in proper orders. Subsequently, proper combination objective function is selected to judge whether two adjacent sides shall be combined. To ensure the accuracy, all sides are judged. Finally, several sides are combined gradually until get the appropriate clustering result.

Although it has been improved a lot compared to classical hierarchical clustering algorithms, it still has some shortcomings:

1) Low efficiency. Since the time-consuming Prim or Kruskal algorithm is the main algorithm used to build MST, the time complexity of hierarchical clustering algorithm based on MST is the function of n^2: $0(n^2)$. It is almost inapplicable to clustering of big dataset.

2) Great fluctuation of clustering results. It often uses appropriate objective function for side combination, which requires users to set many experimental parameters. As a result, different users with different experiences will come to significantly different clustering results [4].

2.3 DENSITY CLUSTERING ALGORITHM BASED ON MST

In many division approaches, users often groups objects according to their distances. This often ends with a ball cluster. To get different shaped clusters, researchers developed a clustering algorithm based on density. In density clustering algorithm, users preset a threshold. If density of adjacent areas is higher than the threshold, data points will be clustered continuously. In other words, given a class, there must be data points higher or equal to a numerical value in the analysing area. Density clustering algorithm can process “noise” data and get clusters of different shapes. However, it has low efficiency. Its time complexity is generally a function of n^2: $0(n^2)$. Moreover, it performs unsatisfying to dataset with uneven density [5].

Since parameter setting in DBSCAN algorithm is very complicated, Ankerst proposed the OPTICS (Ordering Points to Identify the Clustering Structure) algorithm based on the ordering of classes. It shows good performances to dataset with uneven density. In China, Zhao Yanchang et al. developed an isodense clustering algorithm. In most clustering algorithms based on density, parameter setting affects clustering structure directly and causes violent fluctuation of clustering results.

Analysis: Firstly, users build the MST of given dataset. Secondly, users divide the MST into many subtrees in view of its characteristics. Thirdly, these subtrees will be clustered according to densities of these subtrees, getting corresponding clustering structure.

Principle: Firstly, users analyse feature of given dataset and build the MST. Secondly, they find out the longest $k-1$ sides in the MST and delete them to get an initial cluster valued *k*. The MST after division can be expressed by $T[k]$. Viewed from the space perspective of dataset, the dataset is divided into *k* local areas. Data objects with similar density are divided into the same area. Thirdly, an appropriate function is selected to calculate the function value of each initial cluster. The core object (*d*) of $T[k]$ is determined. All objects within the direct density of d will be clustered. Now, the density clustering result based on MST is acquired.

Such improved density clustering algorithm is significantly superior to the classical ones. However, it still has some shortcomings because of its intrinsic properties.

1) Low efficiency. Since the time-consuming Prim or Kruskal algorithm is the main algorithm used to build MST, the time complexity of density clustering algorithm based on MST is the function of n^2: $0(n^2)$. Its application to clustering of big dataset is restricted.

2) Great fluctuation of clustering results. Users have to set parameters of $T[k]$ when judging objects within the direct density of d. These parameters often have no fixed reference standard and shall be set according to users' experiences. As a result, different users with different experiences will come to significantly different clustering results.

2.4 GRID CLUSTERING ALGORITHM BASED ON MST

Grid clustering algorithm is to develop a network structure by quantize an object space into limited units. It is advantageous for quick operation, but disadvantageous for low clustering accuracy.

To further accelerate the operation of grid clustering algorithm, Wang et al. put forward grid-based multiresolution clustering algorithm. It is quick in operation and has higher data processing efficiency. Moreover, it can make real-time data processing upon data adding and updating in the dataset and produce new clustering result. Though it has evident advantages, it is inferior in clustering accuracy.

Schikuta E. proposed two improved grid clustering algorithms which achieves outstanding performances in clustering of big dataset. These grid clustering algorithms are superior for no consideration to data input sequence, low requirement on data distribution and flexible data dimension and size, but inferior for low clustering accuracy [6].

Analysis: Firstly, users divide the dataset space into many basic grids and then distribute data objects to different grids. Based on users' density threshold, density of grids will be calculated and served as the basis for clustering. Secondly, users can build the MST of dense grids through appropriate algorithm and get *k* clusters by deleting longest *k*-1sides.

Although grid clustering algorithm based on MST has obvious advantages than other classical ones, it is inferior for fluctuating clustering results. As the basis of grid clustering algorithm based on MST, dense grid often needs various parameters to build the MST. Since these parameters often have no fixed reference standard, different users will set different parameters, thus getting different dense grid structures. As a result, they will build different MSTs and finally achieve different clustering results [7].

3 Application of clustering algorithms

Cluster analysis plays an important role in our daily lives. Clustering algorithms, an important mean of cluster analysis, are widely used. For example, they are often used in pattern recognition, research of market prospect, image processing, document classification, etc. [8].

In market analysis, cluster analysis is useful when marketers want to implement different marketing strategies to different customer groups. It can divide customers into groups according to purchase mode and make corresponding marketing strategies to win high market acceptance of their products. Cluster analysis is also highly appreciated in urban planning. It helps analysers and designers to divide and design the region into different types of residential area [9]. In seismic study, researchers can make a cluster analysis on geological faults and divide existing seismic centre into different clusters, so that analysers can get a comprehensive understanding on the distribution of seismic belt. In biological field, cluster analysis on genes of animals and plants enables scientists to classify genes with similar functions and discover deeper information. In real estate sales, salesmen can make different marketing strategies according to the cluster analysis of different commercial residential building [10].

The application of cluster analysis based on MST in genetic noise reduction is introduced. Figure 4 is the model of collected genetic data points, in which there are six noises. Noises in the model were processed using a clustering algorithm based on MST under different *k* and *q*. Results are listed in Table 1. When $k=3$ and $q=2$ or 3, the cluster analysis based on MST achieved the best noise reduction effect.

To verify reasonability of parameter settings, a proof test was conducted. DNA data were acquired using probing tools and 7 elements of the probe data interface $(a_0, a_1, a_2, a_3, a_4, a_5, a_6)$ were collected. Subsequently, the rotation matrix R Equation (5) and the displacement matrix Equation (6) were used to process data. Genetic data points simulated according to Equation (7) are drawn onto a coordinate system. Finally, noises of data points were reduced under the optimal k and q. The test results find good accordance with the cluster analysis.

$$R = \begin{vmatrix} a_0^2 + a_1^2 - a_2^2 - a_3^2 & 2(a_1a_2 - a_0a_3) & 2(a_1a_3 + a_0a_3) \\ 2(a_1a_2 + a_0a_3) & a_0^2 + a_2^2 - a_1^2 - a_3^2 & 2(a_2a_3 - a_0a_1) \\ 2(a_1a_3 - a_0a_2) & 2(a_2a_3 + a_0a_1) & a_0^2 + a_3^2 - a_1^2 - a_2^2 \end{vmatrix}, \quad (5)$$

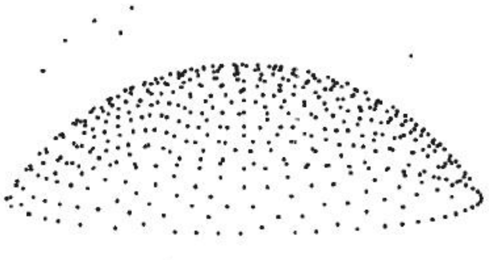

FIGURE 4 Model of genetic data points

$$T = (a_4a_5a_6)^T, \quad (6)$$

$$X = R_1^{-1}(R_2X_2 + T_2 - T_1). \quad (7)$$

TABLE 1 Noise reduction results under different k and q

q	k=1		k=2		k=3		k=4		k=5	
	Noises reduced	**Errors**	**Noises reduced**	**Errors**	**Noises reduced**	**Errors**	**Noises reduced**	**Errors**	**Noises reduced**	**Errors**
1	3	9	3	8	6	2	6	4	6	6
2	3	10	4	8	6	0	6	2	6	7
3	2	6	4	5	6	0	6	5	6	7
4	2	7	5	5	6	0	6	7	6	9
5	3	8	6	4	6	2	6	7	6	10

4 Conclusions

This paper analyses four clustering algorithms based on MST: partitioning clustering algorithm based on MST, hierarchical clustering algorithm based on MST, density clustering algorithm based on MST and grid clustering algorithm based on MST. Principles, shortcomings and applications of these four clustering algorithms are introduced.

Acknowledgements

A Project Supported by Scientific Research Fund of Zhejiang Provincial Education Department (Y201119684).

References

[1] Zhou Y, Grygorash O, Hain T F 2011 Clustering with minimum spanning trees *International Journal on Artificial Intelligence Tools* **20**(1) 139-77

[2] *Deleted by CMNT Editor*

[3] Torkestani J A, Meybodi M R 2011 Learning automata-based algorithms for solving stochastic minimum spanning tree problem *Applied Soft Computing* **11**(6) 4064-77

[4] Yildirim A A, Özdoğan C 2011 Parallel WaveCluster: A liner scaling parallel clustering algorithms implementation with application to very large datasets *Journal of Parallel and Distributed Computing* **71**(1) 955-62

[5] *Deleted by CMNT Editor*

[6] Reddy D, Jana P K 2012 Initialization for K-means Clustering using Voronoi Diagram *Procedia Technology* **4** 395-400

[7] Zhang R, Kabadi S N, Punnen A P 2011 The minimum spanning tree problem with conflict constraints and its variations *Discrete Optimization* **8**(2) 191-205

[8] *Deleted by CMNT Editor*

[9] Tang D 2010 *Research on clustering analysis and its application* PhD thesis of University of Electronic Science and Technology 54-8

[10] Wang X, Liu Q, Lu C 2009 Minimum spanning tree clustering algorithm *Journal of Chinese Computer Systems* **30**(5) 577-822 *(in Chinese)*

Author

Chen Ye, born in February, 1974, Ningbo, Zhejiang Province, P.R. China

Current position, grades: lecturer of the College of Science and Technology, Ningbo University, Zhejiang Province, China.
Scientific interests: statistical pattern recognition and stochastic processes.
Publications: more than 10 papers.
Experience: Teaching experience of 14 years, 4 research projects.