Technique of Cluster analysis in Data mining

WenYang Yu [1, a], YuBing Yang[1, b] and XianWei Wu [1, c]

[1]School of Information Engineering, NingBo DaHongYing University, NingBo 315000,China;

[a] seayuweya@126.com, [b] 79632032@qq.com, [c] wxw786@qq.com

Keywords: Data Mining, Cluster analysis, Cluster algorithm.

Abstract. This paper analyse the cluster analysis method and representation cluster algorithm in the area of the Data Mining, and compare the algorithm capability. And also expatiate the application of the cluster analysis in Data Mining.

Introduction

Data mining is the fundamental statistics, statistical methods in the multivariate data analysis, one of the three methods of cluster analysis is the core technologies used in data mining, to become a very active domain in the study of research subject. Clustering analysis based on the simple idea "birds of a feather flock together", according to the characteristics of the things on the clustering and classification. In this paper, the clustering analysis method in the field of data mining and on behalf of the algorithm is analyzed, and the surface on the performance of the commonly used algorithms from several aspects carries on the analysis comparison. Finally expounds the application of clustering analysis in data mining.

1 The classification of the clustering algorithm in data mining field

Clustering algorithms can be roughly divided into the following categories: partitioning methods, hierarchical methods, the method based on density, based on grid method and the method based on the model.

1.1 Partitioning Method

Given a contain n data object or a tuple of database, a partition method of building data c division, each division said a cluster, and c≤n. or less often adopt a divided guidelines (often referred to as the similarity function), such as distance, so that the object is in the same cluster "similar", objects in different clusters is "different". The clustering method to found in the database of small and medium-sized globular cluster. In order to data sets of large-scale clustering, and clustering, dealing with the complex shape method based on classification need to be further extended.

1.2 Hierarchical Method

Hierarchy method for a given level of decomposition of data collection of objects. According to the hierarchical decomposition is a bottom-up or top-down, hierarchical clustering method can be further divided into condensed and divided. Hierarchical clustering method flaw is that once a step (merge or split) is completed, it cannot be withdrawn, so can't correct the wrong decision. Improve the method of hierarchical clustering quality is a promising direction to hierarchical clustering and other technology integration, form a multi-stage clustering.

1.3 Density-Based Method

Is proposed based on density clustering method is to find the clustering results of any shape. The main idea is: as long as the density of the adjacent area of more than a certain threshold, keep clustering. This method can be used to filter "noise" outlier data, discover clusters of arbitrary shape.

1.4 Grid-Based Method

Grid-based clustering method adopts a multi-resolution grid data structure. The object space quantization is a limited number of units, formed a grid structure. All cluster operations are conducted on the grid structure. The main advantages of this method is its processing speed quickly, the processing time is independent of the number of data objects, with only quantify the number of units of each dimension in the space.

1.5 Model-Based Method

Method based on model for each cluster assume a model and to find the best fitting data for a given model. Based on the algorithm of model possibility by building reflects the density function of the spatial distribution of the data points to locate the cluster. The cluster method tries to optimize the given data and some mathematical model between the adaptability.

2.Commonly used clustering algorithm in data mining field

2.1 CLARANS Algorithm（Random search clustering algorithm）

Partition method first suggested some of the most algorithm is very effective to small data set, but no good scalability to large data sets, such as PAM. CLARA is based on the center of C - type algorithm, can handle larger data set. CLARA algorithm does not consider the whole data set, but randomly select a small part of the actual data as sample, and then choose from sample center in PAM way. This is likely to be the center of the pick and choose of the whole data set is very approximate. Repeat this method, finally to return to the best clustering results as output.

2.2 CURE Algorithm（Clustering using Representative）

CURE algorithm based on centroid and based on the representative object methods between the middle of the strategy. The algorithm first look at each data point as a cluster, and then to a specific contraction factor to center their "shrink", namely the merger of the two nearest representative points in the cluster. It avoided to use all the points or a single center of mass to represent a variety of traditional methods, a cluster with more representative points, make the CURE can be adapted to the spherical geometry. In addition, the shrinkage factor decreasing the influence of noise on the clustering, so that the CURE for the processing of isolated points more robust, and can identify the spherical and larger cluster size change. The complexity of the CURE is O (n), n is the number of objects.

2.3 BIRCH Algorithm (Using the method of hierarchical balance iteration reduction and clustering)

BIRCH is an integrated hierarchical clustering method. It USES clustering characteristics and characteristics of clustering tree (CF) to generalize clustering description. Description is as follows:

With N d data points for a bunch of { $\vec{x}_i$ } (i = 1, 2, 3,..., N), the clustering feature vector is defined as:

$$CF = (N\ ,\ \overset{r}{LS}\ ,\ SS) \qquad (1)$$

Where N is the number of cluster midpoint; $\vec{LS}$ for linear and N ($\sum_{i=1}^{N}\vec{o}_i$), reflected the cluster's center of gravity, SS is the sum of the squares of the other data points ($\sum_{i=1}^{N}\vec{o}_i^{\,2}$), reflects the diameters of the class.

2.4 DBSCAN Algorithm (Density clustering method based on the high-density connection area)

DBSCAN algorithm is sufficiently high density area can be divided into clusters, and can be found in the spatial database with "noise" of any shape clustering. This algorithm defines cluster density connected to the maximum set point.

DBSCAN by checking the database ε-neighborhood of each point p to look for the cluster. If a point p’ ε-neighborhood contains more than MinPts, create a new cluster. As a core object and then repeatedly find density can be up to the object directly from these core objects, when there is no new points can be added to any cluster, the process ends. Not included in any objects in the cluster is considered to be "noise". If the spatial index, the computational complexity of DBSCAN is$O(n \log n)$.

2.5 STING Algorithm (Statistical information style)

STING (Statistaical Information Grid_based method) is a kind of multi-resolution clustering technique based on style, it could be divided into rectangular element space area. In view of the different levels of resolution, there are usually several levels of rectangular cell, these units form a hierarchy: at the top of each unit is divided into multiple low layer unit. Statistical parameters of high-level unit can be easily calculated from the lower unit. These parameters include: property unrelated parameters count;Attributes related parameter m (average), s (standard deviation), min (minimum), Max (maximum), as well as the unit of distribution of the attribute values to follow (distribution) type.

2.6 COBWEB Algorithm (Popular simple concept of incremental clustering algorithm)

Conceptual clustering is one of machine learning methods, most of the concept of clustering method adopted the way of statistics, in determining the concept or the use of clustering probability measure. For in the form of a classification tree to create hierarchical clustering, its input object classification attribute pairs is used to describe.

A hierarchy in the classification tree brother nodes form a partition. For using a heuristic estimate measurement, classification utility to guide the building of the tree. Classification utility are defined as follows:

$$\frac{\sum_{k=1}^{n} P(C_k)\,[\sum_i \sum_j P(A_i = V_{ij} \mid C_k)^2 \quad \sum_i \sum_j P(A_i = V_{ij})^2]}{n} \tag{2}$$

n is at some level in tree form a partition node $\{C_1, C_2, \cdots, C_n\}$, concept or "categories".

2.7 Fuzzy clustering Algorithm

The above introduced several clustering algorithms can be exported to determine clustering, that is, a data point or belongs to a class, or do not belong to a class, and there is no overlap. We can call these clustering methods sort "certainty". In some cases, no sure support clustering can introduce the concept of Fuzzy logic. For Fuzzy sets, a data point is based on a certain degree belongs to a class, can also belong to a few classes in different degree.

FCM algorithm, using membership function defined clustering loss function can be written as follows:

$$J_f = \sum_{j=1}^{c} \sum_{i=1}^{n} [\mu_j(x_i)]^b \,||\, x_i \quad m_j \,||^2. \qquad （3）$$

3. Compare the performance of clustering algorithms

Based on the above analysis, the following the performance of the commonly used clustering algorithm from the scalability, found that the shape of the clustering, sensitivity to the "noise", the sensitivity of the sequence of data input, high dimension and six aspects to compare the algorithm efficiency, as shown in table 1.

Table 1 Compare Clustering Algorithm

	scalability	Found that the shape of the cluster	Sensitivity to the "noise"	Sensitivity to the data input sequence	High-dimensional sex	The algorithm efficiency
CLARANS	good	Convex or spherical	Not sensitive	Very sensitive	general	The lower
CURE	poor	Arbitrary shape	Not sensitive	sensitive	good	higher
BIRCH	poor	Convex or spherical	general	Don't too sensitive	good	high
STING	good	Arbitrary shape	Not sensitive	Not sensitive	good	high
DBSCAN	better	Arbitrary shape	Not sensitive	sensitive	general	general
COBWEB	better	Arbitrary shape	general	sensitive	good	The lower
FCM	good	Arbitrary shape	sensitive	Not sensitive	good	higher

As a result of data mining in different areas of the application of clustering algorithm is put forward their own special requirements, table 1 can give clustering algorithm for the selection of the research and application of reference.

4. The application of clustering analysis in data mining

Clustering analysis in the application of data mining mainly has two aspects: one, the cluster analysis can be used as a preprocessing step of other algorithm, the algorithm is then processed in the generated clusters. Can be used as a preprocessing step, feature and classification algorithm and clustering result can be used for further correlation analysis. 2, can be used as a stand-alone tool for data distribution situation, observing the characteristics of each cluster, some clusters of specific focus for further analysis. Available in the market segmentation, target customer orientation, performance appraisal, biota division, etc., such as in business, clustering analysis can help the market analysts found that different customer base, from basic library client and buying patterns to describe the characteristics of the different customer groups. Three, clustering analysis can complete outlier mining. Many data mining algorithm attempts to minimize outlier effects, or eliminate them. However, isolated point itself may be very useful. As in fraud detection, outlier could herald a fraud.

References

[1] M.Halkidi, Y.Batistakis, M.Vazirgiannis Clustering algorithms and validity measures IEEE 2011.3-22

[2] GEHRKE J,AGRAWAL R,GUNOPULOS D. Automatic Subspace Clustering of High Dimensional Data fro Data caitons[J]. ACM SIMOD,2012,72(2):94-105.

[3] Ng RT, CALBERSON J,Efficient and Effective Clustering Methods for Spatial Data Mining[A].In:Porc of the V[C].Santiago,Chile,2014.144-155.