Research on Improved Partition Clustering Method Based on K-Means Algorithm

Qing Tan[a,*], and Wuchao Zhao

Luoyang Normal University, Henan Luoyang, 471934, China

[a]edutanqing@163.com

Keywords: K-Means Algorithm; Partition Clustering; Mountain Climbing; Repeated Loop Structure; V Data Mining

Abstract: Partition-based clustering algorithm is an optimization search algorithm based on mountain climbing, which is simple, fast and effective. The criterion of dividing is that the data objects in the same cluster are as similar as possible, and the data objects in different clusters are as different as possible. The k-means algorithm is a classical algorithm to solve the clustering problem. The most important feature of the algorithm is that it adopts a two-stage repeated loop structure. The condition of the end of the algorithm is that no data elements are redistributed. The paper presents improved partition clustering method based on K-means algorithm.

1. Introduction

In the stage of data mining, according to the definition of business problem in the first step, the task or purpose of mining is defined, such as classification, clustering, association rule discovery or sequential pattern discovery, and the appropriate data mining algorithm is selected.

In real life, there is often a problem: the collection of the data object set to be analyzed is not numbered. When analyzing such data, it is necessary to use a poly [1]. a clustering method is a process of forming a collection of physical or abstract objects into a plurality of classes made up of similar objects. A cluster consisting of a poly is a collection of a set of data objects that are similar to one another in the same cluster, distinct from objects in other clusters.

The k-average algorithm takes k as the parameter and divides n objects into k clusters in order to make the cluster have higher similarity and lower similarity among clusters. The similarity is calculated according to the average value of the objects in a cluster (considered as the center of gravity of the cluster).

Correlation clustering algorithm based on density is a common clustering algorithm based on density. This algorithm first needs to set any data object as the core data object. The number of data objects contained in the scope of Eps is not less than the number specified by Minpts, and then the core objects are merged according to the corresponding rules, and the clustering analysis of clusters is completed.

Hierarchical clustering consists of different levels of segmentation clustering. It does not require input parameters, which is an obvious advantage of it over the segmentation clustering algorithm, and its disadvantage is that the termination condition must be specified concretely [2]. Typical hierarchical clustering algorithms include BIRCH algorithm, DBSCAN algorithm and CURE algorithm.

Different clustering algorithms are formed by different judgment of similarity. There are many kinds of clustering algorithms, which can be divided into traditional clustering algorithm and modern clustering algorithm in time, soft clustering and hard clustering from subset elements. The initial state can be divided into structural clustering and decentralized clustering, and there are other classification methods based on other criteria.

Groups cases based on a set of attributes. The case in the same cluster has more or less the same attribute value. Clustering is a kind of unsupervised data mining task. There is not a single attribute to

DOI: 10.25236/isrme.2019.006

guide the construction of the model, and all attributes are treated equally. This paper will discuss the clustering analysis technology in data mining.

Clustering can be used as an independent tool to obtain the distribution of data, to obtain useful summary and interpretation of the data in the problem; Clustering can be used as a preprocessing step of other algorithms such as initiating a supervised learning statistical classification method to provide centroid estimation for neural network classifiers.

2. Working Principle Analysis of K-Means Clustering Algorithm

The selection of a large number of clustering algorithms depends on the type of data, the purpose and application of clustering. If a cluster analysis tool is used for description or exploration, multiple algorithms can be tried for the same data to discover the results that may be revealed by the data.

Hierarchical correlation clustering algorithm is a simple and basic hierarchical clustering algorithm [3]. The algorithm has good clustering performance. It mainly includes two concepts: clustering feature (CF) and clustering feature tree (CF-Tree). These two concepts are used to describe and enable the algorithm to deal with data sets effectively.

The algorithm firstly selects k points from the data set randomly as the initial clustering center, then calculates the distance between each sample and the cluster, and classifies the sample to the cluster center nearest to it. The average value of each newly formed cluster data object is calculated to obtain the new clustering center.

Random search clustering algorithm is a segmentation clustering method. First, it selects a point randomly as the current point, then randomly checks some adjacent points around it that do not exceed the parameter Maxneighbor number. If it finds a better adjacent point than it, then it is transferred to the adjacent point, otherwise, the point is regarded as the local minimum. Then randomly select one point to find another local minimum.

Most of the clustering algorithms belong to hard clustering, each element can only belong to one set, and the clustering result will be affected when the feature of the element is blurred. In addition, some traditional clustering algorithms need to input the initial value of the number of subsets, which makes the clustering effect is not good when dealing with big data, as is shown by equation(1) [4].

$$C(t) = \frac{E[B(t), B(-t)]}{E[B(t)^2]} = 2^{2H-1} \tag{1}$$

The mean vector of the samples contained in each cluster domain is obtained: where C(t) is the number of samples contained in the Sj of the j th clustering domain. Taking the mean vector as the new clustering center, the following clustering criterion functions can be minimized: in this step, the sample mean vectors in K clusters should be calculated separately, so it is called the K-means algorithm.

It shows that the clustering criterion function has converged at the end of the sample adjustment. One of the features of this algorithm is that the classification of each sample is correct in each iteration. If it is not correct, it is necessary to adjust, after all the samples are adjusted, then modify the cluster center to enter the next iteration.

Among them, k represents the total number of n factors in the system to be tested, that is, the sum of the number of factors in the set f, and the number of each value is 1 / 2, … The value of a ~ (1) indicates whether the first discrete value and the j _ (j) discrete value overlay are to be covered, and a ~ (1) indicates that the first discrete value and the j _ (th) discrete value pair need to be overlaid.

The time complexity of approximate spectral clustering algorithm for outlier optimization is analyzed simply [5]. Step 1: the time complexity of constructing similar matrix using Gao Si function formula is, where the number of data points is represented. The time complexity of calculating the similarity between the data points and the data points is the time complexity of the whole data set. Step 2: using the sparse matrix to obtain the positive semidefinite matrix and adjust it to the symmetric positive semidefinite matrix with the help of the maximum heap, its time complexity is, where is the nearest neighbor number.

Given the number of partitions to build, create an initial partition. Then an iterative repositioning technique is used to improve the partition by moving objects between partitions. The general criterion for determining a good partition is that objects in the same class are as close or relevant as possible, and objects in different classes are as far away from or different as possible. In order to achieve global optimality, partition-based algorithms require exhaustive partitioning as much as possible.

The processing flow of K-average algorithm is as follows. First, k objects are randomly selected, each of which initially represents the average or center of a cluster. For each remaining object, it is assigned to the nearest cluster according to its distance from the center of each cluster. The average value of each cluster is then recalculated. This process is repeated until the criterion function converges.

A model-based clustering method is proposed. The algorithm is divided into two steps: expected step and maximum step. Given the current cluster center, each data object is divided into the cluster closest to the cluster center, and then the maximum step is to adjust each cluster center so that the sum of the distance between the dispatched data object and the new center is minimized. Until the clustering converges or changes sufficiently small.

3. Research on Improved Partition Clustering Method Based on K-Means Algorithm

The core of the clustering method is to use a poly-like feature 3-tuple to represent the relevant information of one cluster, so that the representation of a cluster point can be represented by the corresponding poly-like feature without having to be represented by a specific set of points. it seeks the poly by constructing a class of characteristic trees that satisfy the branch factor and the cluster diameter limit [6]. The BIRCH algorithm can be used to calculate the distance between the center, the radius, the diameter and the class.

Clustering analysis is an important function in data mining, and clustering algorithm is the core in the field of clustering mining. The quality of clustering algorithm depends on the criterion of similarity, the realization of the algorithm and the ability of discovering hidden patterns.

The k _ means algorithm is a hard-partitioning criterion so that each object can only be divided into one class[5]. The core idea is to put n objects xj (j = 1,...) under the condition that the nonlinear objective function satisfying the formula (1) is minimized by continuously iterating. and n) is divided into k classes ci (i = 1, 2,...). and k), so that the objects in the class have higher similarity, the similarity of the inter-class objects is low, and the generated class is as compact and independent as possible.

The reason that does not need to be covered is a value pair inside the same factor, Or a combination pair is a reversal of an existing combination pair. The problem of generating a combination test dataset with pairwise coverage is to find the smallest set of test data pairs that cover all factors. In the above model, we find the minimum set of all pairwise pairs covering the binary relation matrix t = (t I, j) × p, so as to reduce the test cost as much as possible under the premise of ensuring the ability of error detection, as is shown by equation(2).

$$MSE = \frac{1}{MN}\sum_{k=1}^{M}\sum_{j=1}^{N}(\widetilde{P}_{x_k y_j} - P_{x_k y_j})^2 \tag{2}$$

The optimization of outliers is a difficult problem with non-deterministic polynomials. The time complexity increases exponentially with the dimension of approximate similarity matrix. Step 4 and step 5: the time complexity of computing symmetric semidefinite Laplace matrix and finding out the eigenvector of k minimum nonzero eigenvalues has been analyzed in detail in the second section of chapter 2, that is; Step 6: the time complexity of calculating the normalized matrix is; step 7: the time complexity of executing k-means clustering is: where the number of iterations in the k-means clustering process is represented by the number of clusters.

The advantages of this algorithm are: (1) simple and fast, (2) scalable and efficient for big data set; (3) the algorithm tries to find k partitions which minimize the squared error function value. When the result clusters are dense and the difference between clusters is obvious, the effect is better.

Suppose there is a set of objects distributed in space. Given KG 3, it is required to cluster these objects into three clusters. According to the k-average algorithm, we choose three objects as the center of the initial cluster, and the center of the cluster is indicated by "+" in the graph. Each object is assigned to the nearest cluster according to the distance from the center of the cluster. This distribution forms the pattern.

Most clustering algorithms often require high space-time overhead when processing high-dimensional data, and the results of the algorithm are often disappointing, and the main reason for this is that two are: On the one hand, the high-dimensional data set may not be clearly described and displayed on the human's thinking ability and the visual sense, the difference in the information contained in the different high-dimensional data sets is very large, and the same algorithm cannot be well adapted to the actual high-dimensional data set.

The weak dependence of input parameters on domain knowledge: in clustering analysis, many clustering algorithms require users to input some parameters, such as the number of clusters to be discovered [7]. Clustering results are usually very sensitive to user input of these parameters, and for high-dimensional data, these parameters are sometimes very difficult to determine. This not only increases the burden of users, but also makes the quality of clustering difficult to ensure.

The improvement of the algorithm is as follows: (1) the k- norm algorithm can calculate the discrete attributes; (2) instead of using the average value in the cluster as the reference point, the object located closest to the center in the cluster is chosen, that is, the PAM algorithm.

4. System Experiments and Analysis

Aggregation clustering initially considers all items to be clustered as an independent class, and uses join rules, including single link, full join, average join between classes, and Euclidean distance as similarity calculation algorithm. Combine the two classes with the highest similarity into one class. This is repeated until all items are merged into the same class.

The clustering feature tree of the algorithm is a highly balanced tree with two parameter branching factor B and the class diameter T. The branching factor specifies the maximum number of children in each node of the tree, and the class diameter reflects the restriction on the diameter of a class of points, that is, in what range these points can be clustered into a class, and the non-leaf node is the largest keyword for its children. You can index people based on these keywords, which summarize the information of their children.

The usual method is to select k data objects randomly as the initial clustering center points, and then use an iterative relocation technique to try to improve the partition by moving the objects between the partitions. Each improved packet scheme is better than the previous one.

The K-means clustering method can be divided into the following steps: first, initialize the cluster center 1, according to the specific problems, select C samples from the sample set as the initial cluster center. Using pre-C samples as the initial clustering center. 3. All samples are randomly divided into class C, the mean values of each class are calculated, and the mean values of samples are taken as the initial clustering centers. Second, initial cluster 1, according to the principle of proximity, the sample is classified into the cluster represented by each cluster center, as is shown by equation(3).

$$R_f(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) f(t-\tau) dt \tag{3}$$

The K-average method is used only if the average value of the cluster is defined. This may not be suitable for some applications. For example, data involving classified attributes. The requirement that the user must give the number of clusters to be generated in advance may be a disadvantage of this method.

Attribute values are expressed in real numbers, while typical examples include weight, height, and temperature. In order to classify the data sets, we must define the measure of heterogeneity and similarity to measure the similarity of the data of the same category and the difference between the data of different categories. Clustering analysis can be affected if multiple attributes of the data use different units of measurement.

For a given database containing n data objects, the data is usually divided into k clusters by using the objective function minimization strategy based on the partition method, which requires the user to give the final partition number of the constructed data. Each cluster meets the following two conditions: (1) each cluster contains at least one data object; (2) each data object belongs to one cluster (note: the second requirement can be relaxed in some fuzzy partitioning techniques).

The improved algorithm has four stages. In the first stage, all data items in the cluster are scanned and an CF tree is initialized according to the initial threshold T. In the second stage, the CF tree is reconstructed by increasing the threshold T to increase the degree of aggregation. In the third and fourth stages, the existing CF trees are clustered globally to obtain better clustering results.

5. Summary

Clustering algorithms generally have five methods, the most important are partitioning method and hierarchical method. The partition clustering algorithm divides the data set into K parts by optimizing evaluation function, which needs K as input parameter. Typical segmentation clustering algorithms have K-means algorithm, K-medoids algorithm, CLARANS algorithm. On the basis of sample similarity, clustering analysis requires a certain criterion function to aggregate the samples belonging to the same class into one class and separate the samples from different classes. If the clustering criteria are chosen well, the clustering quality will be high.

References

[1] K.Koperski, J.Han, Discovery of spatial association rules in geographi information databases, in: Proc. of internationgal Symposium on Advance in Spatial Databases, SSD, LNCS, vol.951, Springer Verlag, 2015:47-66.

[2] zalik k r. an efficient k_means clustering algorithm.pattern recognition letters, 2018, 29(9): 1385-1391.

[3] T. Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Record. Vol. 25. No. 2. ACM, 2016.

[4] Falkenauer E, The grouping genetic algorithms, Widening the scope of the Gas, Belglan Joural of Operation Research Statistics And Computer Seinece, vol.33, 2013:79~102.

[5] Hamerly G, Elkan C. Alternatives to the k-means algorithm that find better clusterings. ACM, 2012(11):600-607.

[6] T Zhang, Rramakrishnan, Mogihara.An efficient data clustering method for very large databases. In Pror.2016.ACM-SIGMOD Int. Conf. Management of Data, Montreal, Canada, June, 2016.103-114.

[7] Tsai M. Du, C.C. Lin. Fuzzy C-means based clustering for linearly and nonlinearly separable data. Pattern Recognition, 2011, 44: 1750~1760.