Research and Improvement of Clustering Algorithm in Data Mining

Ren Jingbiao,Yin Shaohong
Tianjin Polytechnic University
rjs123@126.com

***Abstract*-this paper is a cluster analysis algorithm research carried out based on the existing data mining, which focuses on the current popular and commonly used K-means algorithm, and presents an improved K-harmonic means clustering algorithm through using a new distance measure. Through the regulation of distance metric parameters can achieve better clustering effects than the traditional K-harmonic means, and has an advantage both in run time and number of iterations.**

Keyword-data mining; clustering analysis; K-means algorithm

I. Introduction

With the development of computer science and technology, computer has been applied to all walks of life, and the efficiency, effectiveness computer software brought has been continuously attended. Especially in recent years, with the continuous development of the database technology and information technology, people's capacity to produce and collect data has rapidly increased, and the size of the database has expanded rapidly. Whether commercial enterprises, research institutions or government departments, in the past several years have accumulated massive amounts of data stored in different forms, because these data are very complicated, so how to get valuable information or knowledge from them to achieve the purpose for decision-making becomes a very difficult task[1]. As a result, data mining techniques have emerged, and showed an unprecedented strong vitality, and thus gradually become a research hot spot, attracting a lot of people to study. And clustering analysis is a very important data mining technology and method, is one of the main tasks of data mining.

II. Cluster Analysis in Data Mining

As an important branch and effective tool of data mining, cluster analysis is not a new area, which has already been applied to other disciplines.

A. Data mining description

Data mining is a process to extract the implicit, not known in advance and potentially useful information and knowledge from a large number of incomplete, noisy, vague and random practical application data.

Data mining is a reliance on the application, different data mining applications may require different data mining techniques, and the processing flow may also various, the general data mining process as shown in figure 1.

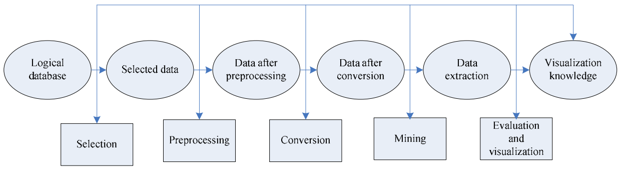

Figure 1. Basic process of data mining

B. Clustering analysis overview

Clustering analysis is a data mining technology to divide the data objects into more than one classes or clusters, and in data mining the clustering analysis focuses on the scalability of clustering methods, data validation for clustering complex shapes and complex types[2], high-dimensional clustering analysis techniques, as well as for the mixed values of large databases and classification data clustering methods.

C. Clustering general steps

In practical clustering analysis, according to whether the domain knowledge involved in the whole process can be broken down into three links, each link has its clear task, so that the whole process of clustering analysis can be clearly understood, as shown in figure 2.

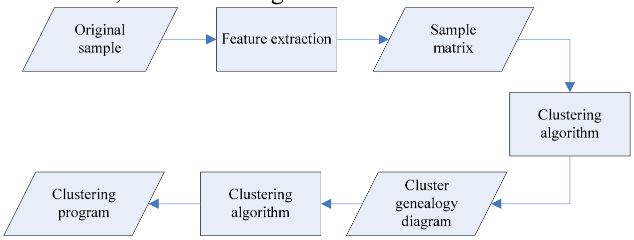

Figure 2. General steps of clustering

1) Feature extraction

Its input is the original samples used by the domain experts to decide what characteristics used to deeply characterize the nature and structure of the sample. Its result output is a matrix, and each row of which is a sample, each list is a characteristic indicator variable.

2) Implementation of clustering algorithm, access to cluster genealogy diagram

Its input is a sample matrix, which is to think a sample to be a point in the characteristic variable space. The output of the clustering algorithm is usually a cluster genealogy diagram to reflect all the classification; or simply give a specific classification scheme, including total categories, and

978-1-4244-6893-5/$26.00 © 2010 IEEE

each category contains what sample points, and so on.

3) Select appropriate classification threshold

After obtaining a cluster genealogy diagram, the domain experts will decide the threshold selection according to the specific applications by experience and domain knowledge. After the threshold selection the classification schemes can be directly seen from the cluster genealogy diagram. Experts in the field can also further analyze the cluster results with domain knowledge, so that to deepen the sample points and characteristic variable understanding.

D. Main algorithm for clustering analysis

An important step of the data mining is data preparation, which includes standardization, integration and pre-processing of the selected data, etc., which is a prerequisite for data mining, also a necessary prerequisite for the normal operation of the clustering algorithm.

There are many clustering algorithms, need to apply the type of data involved, and the purpose of clustering, as well as specific application requirements to select the appropriate clustering algorithm[3]. Generally the clustering analysis algorithms can be divided into the following categories: classification methods, hierarchical methods, density-based methods, grid-based methods and model-based methods.

III. K-MEANS CLUSTERING ALGORITHM

K-means clustering algorithm is one of the commonly used partition clustering algorithms.

A. K-means clustering algorithm principle

K-means algorithm is to divide the n-data objects into K clusters to make the obtained clusters meet the following requirements: the objects in same clustering have higher similarity, while the objects in different clustering have smaller similarity.

K-means algorithm is a dynamic clustering algorithm based on the standard measure function.

K-means algorithm is to divide the n vectors $X_j(j=1,2,...n)$ into K classes $G_i(i=1,2,...k)$, and seek each type of cluster center, making the standard measurable function lower than the given minimum threshold value or twice the value less than a parameter threshold, out of service.

Once selected the criterion function, clustering has become a problem with a clear definition in class discrete optimization problems, making the criterion function take extreme value. Sample set is limited, so the desired division way is limited. In theory, the clustering problem can always be solved through the exhaustion[4]. However, apart from dealing with those very simple problems, the exhaustion has no practicability for its computational complexity. Therefore, the exhaustive method is not applicable for most clustering applications.

In practical applications, using heuristic methods, and using the average of each category to represent the class, which greatly reduces the computational complexity, increases the computing speed, so that large-scale data processing is possible. At the same time, this also leads to the method under a great impact of the initial value, because the objective function has many local minimum points, while the iteration of the algorithm is carried out along the direction of the objective function decrease, if the initialization is dropped in the vicinity of the local minimum point, it will result in algorithms converge to a local minimum.

B. K-means algorithm

In the each step of the iteration in the classical K-means algorithm, each sample point is considered to be completely belonging to a category. The fuzzy K-means algorithm steps are shown in figure 3.

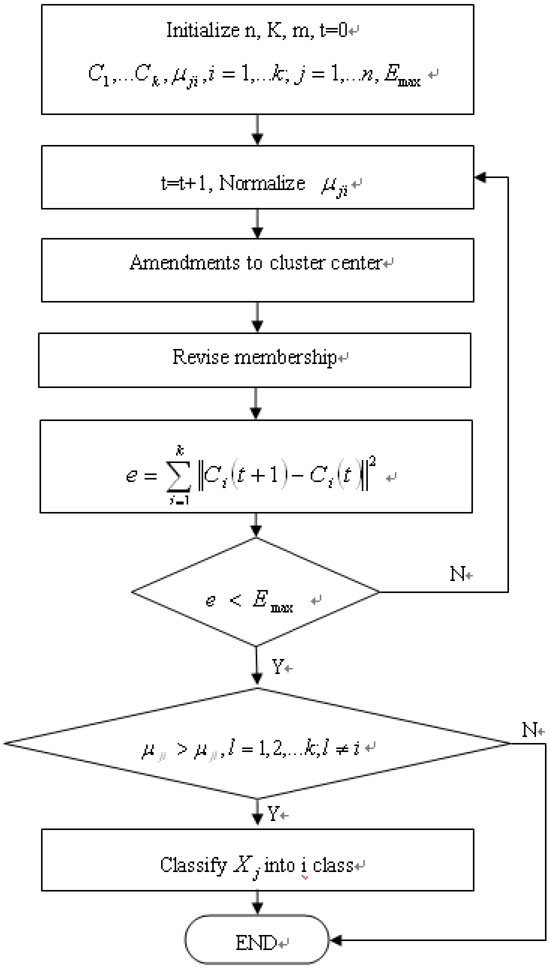
Initialize n, K, m, t=0
$C_1,...C_k, \mu_{ji}, i=1,...k; j=1,...n, E_{\max}$
t=t+1, Normalize μ_{ji}
Amendments to cluster center
Revise membership
$e=\sum_{i=1}^{k}\left\|C_i(t+1)-C_i(t)\right\|^2$
$e < E_{\max}$
N
Y
$\mu_{ji} > \mu_{jl}, l=1,2,...k; l \neq i$
N
Y
Classify X_j into i class
END

Figure 3. Fuzzy K-means algorithm steps

IV. K-Means Algorithm Implementation and Improvement

A. Improved K-Means harmonic algorithm

K-Means algorithm needs to give the clustering number to be built, which will first create an initial division[6], and then use an iterative relocation technique to try to improve the division through an object moving along it.

As the K-harmonic means does not rely on the initial point, clustering time and clustering results, so we can add some improved conditions and factors based on this algorithm, thus to further enhance the effectiveness of this algorithm clustering. This paper will apply a new distance measurement to the K-Harmonic means, and test the effectiveness of this combination through numerical experiments.

The new measurement function is as follows:

$$d(x,y)=1-\exp\left(-\beta\|x-y\|^2\right) \tag{1}$$

In which β is a positive constant, from this distance function it can be seen that $d(x,y)$ is a monotonically increasing function about $\|x-y\|$, namely $d(x,y)$ increases with the increase of $\|x-y\|$.

As we all know, to make a point weight and more robust, it would meet the weight of the abnormal points and noise points to be smaller, and the weight of the compact point with data concentration should be greater. This new measurement is precisely to meet this requirement. Through which, the object function of the improved K-Means clustering algorithm can be obtained as follows:

$$AKM(X,C)=\sum_{j=1}^{k}\sum_{x_i\in C_j}\left\{1-\exp\left(-\beta\left\|x_i-w_j\right\|^2\right)\right\} \tag{2}$$

For the traditional K-means algorithm, the weight of each data point is 0 or 1, so the traditional K-means algorithm also can be called as a hard K-means algorithm. After the application of the new distance measurement, the weight of the data point is longer only for 1 or 0, but determined by the coefficient $\exp\left(-\beta\left\|x_i-w_j\right\|^2\right)$, which will transfer the hard division into soft division[5]. But precisely because of this role, the improved algorithm has better clustering results than the traditional clustering algorithm.

B. Numerical experiment results and analysis

In this paper, IRIS data is used for numerical experiments to different β selections. Table 1 is the obtained clustering results, and the last list is the clustering results for the traditional KHM algorithm.

TABLE I. PARAMETER B EFFECTS TO KHM IN IRIS

B value	Number of clustering categories	Running time/s	Iterations	Error square sum
10	130 9 11	0.2737	15	2.5932
5	110 21 17	0.2675	15	2.2746
1	87 37 19	0.2597	15	2.1397
0.1	50 43 52	0.2369	15	0.0043
KHM	45 32 50	0.2019	18	0.4981

As can be seen from the table, when the β value is not the same, AKHM algorithm iterations are the same, both 15 times, running time or less, while the error square sum is gradually reduced as the β reduced, and the number of clustering types is better. In general, the smaller the error square sum, the more correct this parameter selection. Moreover, according to the preceding analysis, in order to avoid all of the data assigned to the same data, the choice of parameter β should not be too small. At the same time, the error square sum of the KHM algorithm in IRIS data is much greater than the improved KHM algorithm. The number of iterations is 15 times, whereas the traditional KHM algorithm iterations is 18 times, and in the running time, the improved KHM is not very different from the traditional KHM algorithm, running time is fast, from which it can be seen that for the IRIS data, the combination of the new distance measurement and the K-Harmonic means is a very effective and feasible way to improve.

According to analysis and numerical experiments, through a combination of a new distance measurement, AKHM algorithm changed the objective function of the original KHM algorithm, at the same time, because in the new distance measurement there is a adjustable parameter, by adjusting the parameter AKHM algorithm can obtain clustering results better than the KHM algorithm, and has certain advantages both in the run time and the number of iterations.

V. Conclusion

This paper carried out the clustering analysis algorithm research based on the existing data mining, focused on the current popular and commonly used K-means algorithms, and proposed an improved K-Harmonic means clustering algorithm by adopting a new distance measurement. Through the regulation of distance metric parameters can achieve better clustering effects than the traditional K-harmonic means, and has an advantage both in run time and number of iterations.

References

[1]. Cai WL, Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. Pattern Recognition, 2007, 40(3):825-833.

[2]. Brendan J F. Delbert D. Clustering by Passing Messages Between Data Points. Science, 2007, 315:972-976.

[3]. Xiaohui Liu, Gongxian Cheng, Wu J.X., Analyzing outliers cautiously, IEEE Tran, on Knowledge and Data Engineering,2002,14(2):432-437.

[4]. Tianming Hu, Ying Yu, Jinzhi Xiong, etal. Maximum likelihood combination of multiple clusterings,2006, 27:1457-1464.

[5]. Daasch W. Robert, Madge Robert. Variance reduction and outliers: statistical analysis of semiconductor test data [C], Test Conference, 2005, 304-312.

[6]. Pal N R, pal K, Bezdek J C. A Possibility fuzzy C-means clustering algorithm [J], IEEE Trans Fuzzy Systems, 2005, 13(4):517-530P.