Wireless Personal Communications
https://doi.org/10.1007/s11277-019-06709-z

Clustering Algorithm and Its Application in Data Mining

Hailei Zou[1]

Abstract
Clustering analysis is one of the main research directions in data mining. At present, it has gone deep into all fields and made good progress. Aiming at the role of clustering analysis in data mining, a clustering analysis algorithm and its application in data mining are proposed. Through literature comparative analysis method, the basic concepts of cluster analysis are expounded in detail, and the classical algorithms in cluster analysis are discussed. The basic realization process of clustering K-means algorithm is analyzed and an example simulation is carried out. The research shows that this algorithm has strong universality and can be applied to most data analysis sites, providing a theoretical basis for timely detection and analysis of large amounts of data.

Keywords Data mining · Clustering analysis algorithm · K-means algorithm · Applied research

1 Introduction

With the rapid development of the Internet around the world and the ever-changing information technology, the kinds of data which people are using are constantly growing at an explosive rate. A large amount of data is stored in the database and can be applied to government offices, business intelligence, scientific research and project development, but to really use these data is not an easy task [1]. Clustering analysis is one of the most important research fields in data mining. Data analysis can find useful information and is widely used in fields such as market research, data analysis, pattern recognition, image processing, and artificial intelligence and Web document classification [2]. In business intelligence applications, clustering analysis can help data mining staff to analyze customer purchasing patterns, characterize different customer groups; special customers can be discovered from customer information database. In biological applications, cluster analysis can be used to infer species information and classify genes according to their functional relevance so that an understanding of the original structure of the population can be obtained [3]. Clustering analysis also helps to identify areas of use-related use in identifying satellite monitoring databases and grouping houses in cities based on their value, type, and location. Different from other data mining methods, the

✉ Hailei Zou
zouhailei@163.com

[1] School of Science, China Jiliang University, Hangzhou, Zhejiang, China

Published online: 21 August 2019

user does not know the content and category of the dataset before using the clustering algorithm. That is, the cluster analysis does not need to be based on prior knowledge and is an unsupervised machine learning [4].

2 State of the Art

Clustering analysis has been a major topic of data mining research for many years. Among them, clustering analysis based on distance is the main content of scholars' research. K-medoids algorithm, K-means algorithm and other clustering algorithm based on clustering mining tools are widely used in many statistical analysis software or system [5]. In the field of machine learning, machine learning is mainly divided into supervised learning and unsupervised learning. Clustering analysis belongs to the category of unsupervised learning. Classification belongs to the category of supervised learning. Clustering does not depend on training samples Classification information, and classification is depend on the classification of training samples information [6]. In order to achieve a better application, the clustering method is combined with other methods to make up for the shortcomings of the method in data mining, and make the performance of clustering method more superior. Methods often combined with clustering methods include: Ant colony algorithm, genetic algorithm and immune algorithm [7]. At present, the researches on clustering algorithms mainly include the following aspects: the selection of initial cluster centers and the input sequence of data sets on the clustering results. In the field of data mining we can use multiple sets of different initial centers for multiple iterative calculations and choose the best one as the final result, but we can not guarantee that this result is the optimal solution, while multiple iterations consume a lot of time, a lot of uncertainty, so it is very important to select the appropriate initial cluster centers, the efficiency of the algorithm. At present, the main research method is to improve the existing clustering by clustering it better [8].

3 Methodology

3.1 Clustering Analysis of the Basic Concepts

Cluster Analysis is a method of studying individuals based on the characteristics of things themselves, with the purpose of classifying similar things. Its principle is that individuals in the same category have greater similarity, and individuals in different categories have the smallest similarity (that is, the difference is greater) [9]. Clustering analysis has the following characteristics: It is suitable for classification without a priori knowledge; it can handle the classification determined by multiple variables; clustering analysis is an exploratory analysis method; clustering mainly focuses on distance-based clustering analysis.

Assuming the data set contains n data objects, there is a data matrix (Data Matrix)

$$\begin{pmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{pmatrix} \tag{1}$$

where x_{if} represents the fth attribute value of the ith object in the data set. The matrix represents the sum of the attribute records for each object in the dataset.

Continuous variable type is a numeric variable of a certain range. Including height, temperature, weight and so on. Calculate the average of the absolute deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|) \tag{2}$$

Among them:

$$m_f = \frac{1}{n}\left(x_{1f} + x_{2f} + \cdots + x_{nf}\right) \tag{3}$$

The normalized measure is:

$$z_{if} = \frac{x_{if} - m_f}{s_f} \tag{4}$$

The corresponding metric distance formula has the following common forms:
Euclidean distance formula:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \tag{5}$$

Manhattan distance formula:

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n| \tag{6}$$

Mingkosiji distance formula:

$$d(x, y) = (|x_1 - y_1|^q + |x_2 - y_2|^q + \cdots + |x_n - y_n|^q)^{\frac{1}{q}} \tag{7}$$

where q is a positive integer. When $q=1$, it represents the Manhattan distance, When $q=2$, it represents the Euclidean distance. Each attribute is assigned a corresponding weight according to its importance. The weighted Minkowski distance formula is:

$$d(x, y) = (w_1|x_1 - y_1|^q + w_2|x_2 - y_2|^q + \cdots + w_n|x_n - y_n|^q)^{\frac{1}{q}} \tag{8}$$

Among them, the weight coefficient to meet the conditions $w_1 + w_2 + \cdots + w_n = 1$.

Discrete variable type, that is, with a finite number of states. For example, the student's age of birth, job title, stage description and so on. A discrete variable state can be letters, symbols, integers, etc., which is equivalent to give it a special meaning, attention does not mean size, but only state description. Thus, the dissimilarity between two objects, all composed of discrete variables, can be calculated by a simple matching method as follows:

$$d(x, y) = \frac{p - m}{p} \tag{9}$$

where m is the number of attributes that match the attribute values in object x and y; p is the total number of attributes.

Mixed types, that is, the data objects of different types of properties. Assuming that the dataset contains p different types of attributes, the dissimilarity $d(x, y)$ of the objects x and y is defined as:

$$d_{xy} = \frac{\sum_{i=1}^{p} \delta_{xy}^{i} d_{xy}^{i}}{\sum_{i=1}^{p} \delta_{xy}^{i}} \tag{10}$$

If x_i or y_i is missing, the function $\delta_{xy}^{i}=0$, otherwise $\delta_{xy}^{i}=1$. d_{xy}^{i} calculation method is related to the specific type of the ith attribute. If it is a discrete variable: when $x_i = y_i d_{xy}^{i} = 0$, else $d_{xy}^{i}=1$. If it is a continuous variable, $d_{xy}^{i} = \frac{|x_i - y_i|}{d_i}$. Where d_i is the difference between the maximum value and the minimum value of the ith attribute value.

3.2 Cluster Analysis Algorithm

First, feature selection. Features must be chosen appropriately to include as much of the task-related information as possible (Fig. 1). Among the characteristics, the redundant reduction and minimization of information is the main purpose. Second, the similarity measure used to quantitatively measure how two feature vectors are "similar" or "dissimilar". A simple measure such as Euclidean distance is often used to reflect the dissimilarity between two eigenvectors. Third, the clustering algorithm. Having chosen the appropriate similarity measure, this step involves selecting a particular clustering algorithm to reveal the clustering structure in the data set. Fourth, the result verification. Once the result is obtained using the clustering algorithm, its validity needs to be verified. Fifth, the result is judged. In many cases, experts in the field of application must use other experimental data and analysis to determine the clustering results, and finally make the correct conclusions [10].

Given the number of clusters k and the objective function F.

The data object set D is divided into k categories, the objective function to achieve the best. The segmentation algorithm transforms the clustering problem into a combinatorial optimization problem. Starting from a set of initial points, the iterative control strategy is used to optimize the objective function. The most commonly used objective function is:

$$\sum_{i=1}^{n} \min_{j=1}^{k} d(x_i, m_j) \tag{11}$$

where m_j is the center of C_j, or the nearest object in the center. Typical algorithms are:

K-means algorithm, k as a parameter, the object is divided into k categories, so that the class has a high degree of similarity. However, the similarity between classes is low, and the

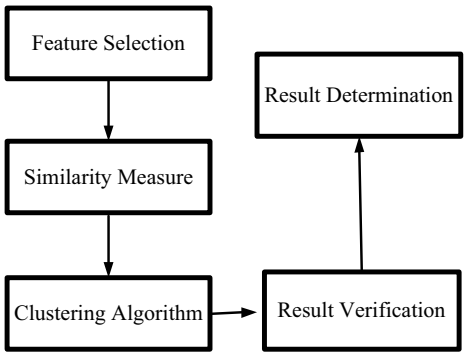

Fig. 1 Clustering algorithm process

similarity calculation is based on the average of the objects in the class. When m_j is the center of C_j in the objective function, such an algorithm is used. The corresponding improved algorithms include the calculation of similarity, the different methods of selecting different average values, and the different strategies of calculating the average of different clusters.

K-center algorithm, each class is free to choose a representative object, and the remaining objects are assigned to the nearest class according to their distance from the representative object, repeatedly using the non-representative object instead of the representative object to improve the clustering quality. When m_j in the objective function is the nearest object to the center in C_j, this type of algorithm is used.

Hierarchical method is the decomposition of a given set of data objects at different levels of data segmentation. With a clear hierarchy, the algorithm can be described by a hierarchical tree. Hierarchical clustering methods are divided into cohesion and decomposition methods, which are based on the level of decomposition is top-down or top-down form. In class merging or splitting, you need to examine the distance between categories. The following methods are widely used to measure the distance between categories:

Shortest distance:

$$d_{\min}(C_i, C_j) = \min_{p} \in C_j \left| p - p' \right| \tag{12}$$

Maximum distance:

$$d_{\max}(C_i, C_j) = \max_{p} \in C_{i,p'} \left| p - p' \right| \tag{13}$$

Average distance:

$$d_{mean}(C_i, C_j) = \left| m_i - m_j \right| \tag{14}$$

The advantage of the hierarchical clustering method is that it can obtain multi-level clustering structures with different granularities, and although it is simple, it often encounters the difficulty of merging or splitting point selection. One promising direction for improving the quality of clustering of hierarchical methods is to integrate hierarchical clustering with other clustering techniques. There are two ways to improve the result of hierarchical clustering: First, a careful analysis of the "joins" between objects in each tier, such as those in CURE and Chameleon, and the second, integrated level cohesion and iterative relocation methods. First the bottom-up level algorithm is used, and then iterative relocation to improve the result is used, for example, the BIRCH method.

3.3 K-means Algorithm Implementation Process

K-means algorithm is a kind of rapid clustering analysis method which is widely used. It has higher execution efficiency and larger sample data volume. However, the sample size of the research design is not large, and the processing time is definitely not the primary consideration in dealing with this type of problem. Therefore, K-means clustering can be considered. It provides a cluster analysis function, which can perform cluster analysis of samples or variables on a variety of data types.

K-means algorithm accepts input k, the n data objects are then divided into k clusters so that the obtained clusters satisfy: The objects in the same cluster have higher similarity, while

the objects in different clusters have less similarity. The clustering result can be represented by a membership matrix:

$$W = \{w_{ij}, 1 \le i \le n, 1 \le j \le k\} \tag{15}$$

$w_{ij} \in \{0, 1\}$, Each object either belongs to a cluster or does not belong, both must live in one. $\sum_{j=1}^{k} w_{ij} = 1$, each column has one and only one element, that is, each object belongs to only one cluster. $\sum_{i=1}^{n} w_{ij} > 0$, each clustering result is not empty.

Any one cluster of D corresponds to one of the above matrices. On the contrary, any one of the matrices satisfying the above conditions also corresponds to a cluster of D. Therefore:

$$M_{hk} = \left\{ W | w_{ij} \in \{0, 1\}, \forall i, j; \sum_{j=1}^{k} w_{ij} = 1, \forall i \in [1, n]; \sum_{i=1}^{n} w_{ij} > 0, \forall j \in [1, k] \right\} \tag{16}$$

Group k clustering space called D. In K-means clustering algorithm, the clustering objective function is:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij} d(x_i, z_j) \tag{17}$$

where x_i is the ith object, z_j is the center of the j the cluster. The purpose of clustering is to find a set of cluster centers and membership matrices that minimize the objective function value.

The realization of the process is as follows.

First, choose k objects randomly from n data objects as the initial cluster center. Second, calculate the distance between each object and these center objects according to the mean (center object) of each clustering object, and then divide the corresponding objects according to the minimum distance. Thirdly, loop from the second step to the third step until each cluster no longer changes.

4 Result Analysis and Discussion

4.1 Experimental Results and Analysis

Clustering analysis of monthly CPI index data uses K-means algorithm and the variable name is defined. Although the regional groupings are not used as analysis variables, they are also entered into the database in order to have a more direct understanding of the clustering results. The variable name is "region". The variable names of the 3-month consumer price index are "January", "February", "March". The growth rate of the consumer price index from January to February and that of the consumer price index from February to March are "growth rate 1" and "growth rate 2", respectively (Fig. 2).

Table 1 is a two-dimensional map of data distribution in each of the 3 months. From which we can get a general understanding of the following situations: the CPI in January was generally high, followed by March, and the consumption in various regions showed a big difference in February. In the meantime, from a general standpoint, the relative changes in consumption in different months among different regions show that there are some

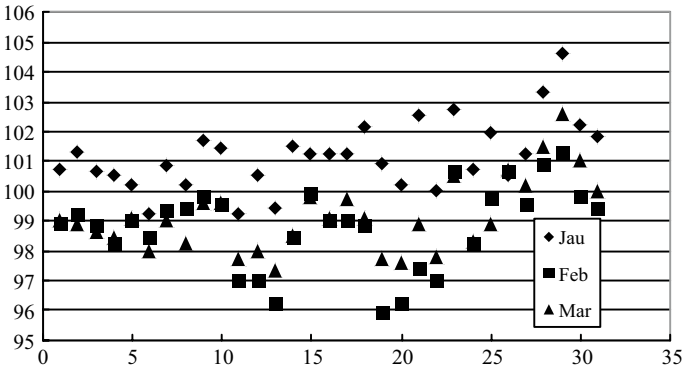

Fig. 2 CPI index scatter plot of each region

Table 1 The object scatter plot with the monthly variable as the sub attribute

	Cluster			
	1	2	3	4
January	104.60	102.20	100.20	99.40
February	101.30	99.80	99.40	96.20
March	102.60	101.00	98.20	97.30

Table 2 Clustering process

Iteration	Change in cluster centers			
	1	2	3	4
1	.000	.956	1.044	.819
2	.875	.234	.073	.000
3	.000	.000	.000	.000

Table 3 Error analysis

	Cluster		Error		F	Sig.
	Mean square	*df*	Mean square	*df*	Mean square	*df*
January	8.641	3	.563	27	15.359	.000
February	16.901	3	.272	27	62.166	.000
March	13.193	3	.226	27	58.461	.000

major fluctuations in some regions and some are relatively stable. This can describe the development of the area from the side and provide a more comprehensive reference value for horizontal comparison.

Four regions randomly selected from 31 regions as the initial clustering centers are not optimally overlapped by the K-means algorithm, and the distances among the classes are

Table 4 Final clustering results

Cluster	
1	2.000
2	10.000
3	13.000
4	6.000
Valid	31.000
Missing	.000

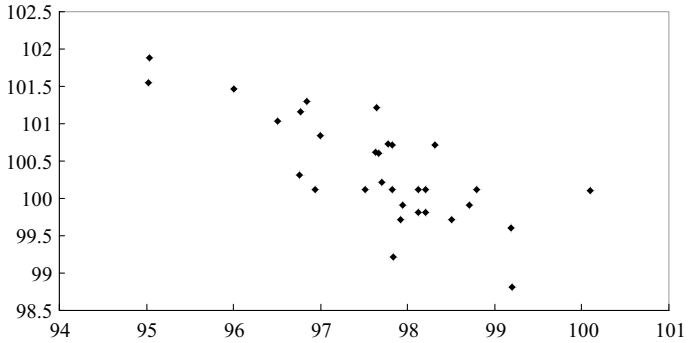

Fig. 3 Growth two-dimensional scatter plot

not optimal. After the iteration, the center values of all the variables in the categories are corrected, as shown in Table 2 shows.

An analysis of variance (ANOVA) was performed on the distance between clusters of the clustering results. Analysis of variance (Table 3) indicated that the probability of difference in distance between categories was < .001. In this way, the original 31 objects are aggregated into 4 categories (Table 4), the first category includes the original two categories, the second category includes the original ten categories, the third category includes the original category 13, the fourth category includes the original six categories. The specific results of the system are a variable stored in the original database.

With reference to professional knowledge, the CPIs of various regions from January to March can be divided into the following four categories: Category 1—reflecting the relatively highest price trend of residents' purchasing of consumer goods and service items during the period, as follows 2 Regions: Gansu and Qinghai; Category 2—Reflecting the relatively high price changes of residents' purchase of consumer goods and services during the period, there are 10 regions as follows: Hubei, Jiangsu, Ningxia, Shandong, Shaanxi, Shanghai, Sichuan Tibet, Xinjiang and Yunnan. Category 3—reflects the relatively low price trend of residents' purchasing of consumer goods and service items in this period. There are 13 regions as follows: Beijing, Guizhou, Hebei, Hainan, Henan, Heilongjiang, Hunan, Jilin, Jiangxi, Liaoning, Inner Mongolia, Shanxi and Tianjin, Class 4—reflect the relatively lowest price trend of residents' purchasing of consumer goods and service items in this period. There are six regions as follows: Anhui, Guangdong, Fujian, Guangxi, Chongqing, Zhejiang.

In order to show the changes of CPI over time, they can be clustered reasonably by the rate of change, and more dynamic information about consumption can be obtained, which helps to know more about the urban development and residents' living

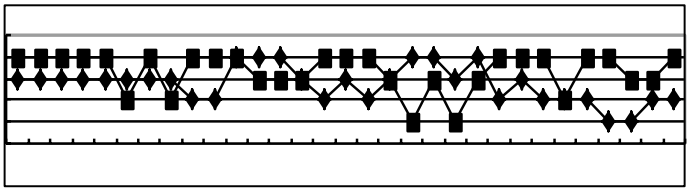

Fig. 4 Comparison of two clustering groups

development. The development of all regions is a very good reference point, and at the same time, it will play a good reference role in the formulation of policy and orientation. Therefore, it will play a good role in promoting the balanced and healthy development in all regions as a whole. The specific experimental results (Figs. 3, 4).

The above figure shows the plot of the clustering conducted according to the January–March consumer price index and the monthly change of the consumer price index for clustering. The following conclusions can be drawn: residents purchasing consumer goods and service items In contrast, the regions with relatively low change trend have lower CPI, on the other hand, the regions with relatively higher price trend of residents purchasing consumer goods and service items have higher CPI. Indicating that the CPI changes in the larger area of the relative residents have to bear greater price volatility and higher service items and out. Which can be identified according to the above map of each region belongs to categories, but also shows the geographical and economic relevance of the degree of development.

5 Conclusion

With the development of society and science and technology, the big data of society has been paid more and more attention by people and the information that people can use is also increasing. However, users' ability to process and understand these data information remains the same. How to accurately find the parts of their interest from these huge data information and how to classify these information involves a new direction, that is, data mining research. The text proposes a method of research and analysis using clustering algorithm in data mining. Using data mining technology, the role of clustering analysis algorithm in information mining is studied in detail, and an example is given to analyze the operation of clustering K-means algorithm. The results show that in the K-means clustering algorithm, we must first determine an initial value to be divided, and then use the algorithm to effectively optimize the initial partition. It is found experimentally that the key and difficult point of clustering K-means algorithm in data mining is the choice of initial clustering center, which will have a great impact on the clustering result. The experimental results show that the clustering K-means algorithm has high accuracy, strong anti-interference and universality, and has great development prospects.

References

1. Camus, P., Mendez, F. J., Medina, R., et al. (2011). Analysis of clustering and selection algorithms for the study of multivariate wave climate. *Coastal Engineering, 58*(6), 453–462.
2. Hencil, P. J. (2014). Fast and economic clustering algorithms in data mining an analytical research. *Nature Reviews Microbiology, 6*(5), 339–348.
3. Rad, A., Naderi, B., & Soltani, M. (2011). Clustering and ranking university majors using data mining and AHP algorithms: A case study in Iran. *Expert Systems with Applications, 38*(1), 755–763.
4. Song, Y. C., Meng, H. D., O'Grady, M. J., et al. (2010). The application of cluster analysis in geophysical data interpretation. *Computational Geosciences, 14*(2), 263–271.
5. Tjortjis, C. (2011). k-ATTRACTORS: A partitional clustering algorithm for numeric data analysis. *Applied Artificial Intelligence, 25*(2), 97–115.
6. Tonello, L., Conway, D. M. E., Marino, G. A., et al. (2015). Data mining-based statistical analysis of biological data uncovers hidden significance: Clustering Hashimoto's thyroiditis patients based on the response of their PBMC with IL-2 and IFN-γ secretion to stimulation with Hsp60. *Cell Stress and Chaperones, 20*(2), 391–395.
7. Zhao, W., Chen, J. J., Roger, P., et al. (2016). A novel procedure on next generation sequencing data analysis using text mining algorithm: BMC. *Bioinformatics, 17*(1), 213.
8. Zhou, Y., Zhou, Y., Luo, Q., et al. (2017). A simplex method-based social spider optimization algorithm for clustering analysis. *Engineering Applications of Artificial Intelligence, 64,* 67–82.
9. Ji, J., Bai, T., Zhou, C., et al. (2013). An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing, 120*(10), 590–596.
10. Cao, H., Jia, L., Si, G., et al. (2013). A Clustering-analysis-based membership functions formation method for fuzzy controller of ball mill pulverizing system. *Journal of Process Control, 23*(1), 34–43.

Hailei Zou Master of Science, Lecturer. Graduated from the Chong-Qing Normal University in 2004. Worked in China JiLiang University. His research interests include statistical analysis and data processing.