Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Overview of Clustering Algorithms

Zichao Wang

School of Computer Science and Technology, North China Electric Power University, Baoding, China

Abstract: Cluster analysis is a multivariate statistical analysis method for quantitative classification of multi-sample data, and it is a typical representative of unsupervised learning in machine learning. Cluster analysis can be divided into Q-type clustering and R-type clustering according to different application samples, and the clustering criteria are derived from the attribute distance of samples, that is, the similarity degree. Clustering algorithms are commonly used in machine learning, data analysis and other fields. Commonly used clustering methods include hierarchical clustering, k-means clustering, and mean shift clustering, each of which has its advantages and disadvantages. Algorithm selection and parameter adjustment need to refer to specific application scenarios.
Keywords: Unsupervised learning; Similarity; Minkowski distance; K-means clustering.

1. INTRODUCTION

Nowadays, the society is entering the era of big data network and intelligence, with massive data coming, and clustering algorithm can help us flexibly extract the valuable information features we want to obtain from massive data. Whether it is machine learning, data mining, clustering algorithm is a common tool. Therefore, the mastery and application of clustering algorithm has become an indispensable tool and help for data analysis in today's society. This paper will focus on clustering algorithms, mainly introducing the concept, classification, common clustering algorithms and applications of clustering.

2. CONCEPT

2.1 Definition

Cluster analysis, also known as group analysis, is a multivariate statistical analysis method for quantitative classification of multiple samples (or indicators).

This paper gives Everitt's definition of clustering in 1974: the entities in a cluster are similar, but the entities in different clusters are not; A class cluster is the convergence of points in test space, and the distance between any two points of the same class cluster is smaller than the distance between any two points of different class clusters; Class clusters can be described as connected regions in multidimensional space containing relatively high density point sets, which are separated from other regions (class clusters) by regions containing relatively low density point sets.

In fact, clustering is an unsupervised classification, and it has no prior knowledge available.

2.2 The difference between clustering and classification

Classification refers to the classification according to category, grade or nature. Classification requires us to know the classification criteria and categories of data features in advance, and then classify data sets according to the classification criteria and data features. In machine learning, we can label the training data, classify the training data, then learn, and finally acquire a certain classification ability, and then use it to classify the unknown data. This learning process of providing training data is usually called supervised learning.

Clustering, relative to classification, is an act of classifying unlabeled data according to the characteristics of data without obtaining classification criteria and categories in advance. Because we didn't get the classification criteria and categories in advance, we don't know how to classify or which categories, but to achieve the purpose of classification, we can only classify according to the similarity of variables, and the specific method is to classify similar data into one category. Therefore, the advantage of clustering is that it can be classified without training data. The disadvantage is that the classification can only be based on similarity, so the number of clusters can not be well grasped, and the clustering results may not be consistent with the expected classification results. In machine learning, this process of completing classification without providing training data is called unsupervised learning.

2.3 The flow of clustering

1) Data preparation: including feature standardization and dimension reduction;

2) Feature selection: select the most effective feature from the initial features and store it in the vector;

3) Feature extraction: form new prominent features by transforming the selected features;

4) Clustering (or grouping): firstly, select a certain distance function of suitable feature type (or construct a new distance function) to measure the proximity, and then perform clustering or grouping;

5) Evaluation of clustering results: it refers to the evaluation of clustering results. There are three main types of evaluation: external effectiveness evaluation, internal effectiveness evaluation and correlation test evaluation.

3. APPLICATION CATEGORIES OF CLUSTER ANALYSIS

3.1 2Criteria for classification

Cluster analysis is a multivariate statistical analysis method for quantitative classification of multiple samples (or indicators). In specific application scenarios, we generally cluster samples or indicators, and classifying samples is called Q-cluster analysis, while classifying indicators is called R-cluster analysis.

3.2 Q type clustering

Similarity measurement of samples

In order to classify samples quantitatively, it is necessary to describe the similarity between samples quantitatively. A sample itself is characterized by multiple indicators, so we classify samples by multiple indicators. The index value of each sample is different, and each sample value has a certain value range. Therefore, we can describe the index value of a sample as variables such as X, Y and Z,

and describe this sample as a point in N-dimensional (n is the number of indexes) space. Naturally, we can describe the similarity of samples by the distance between points of spatial samples.

Let Ω be the set of sample points, and the distance d(x,y) is a function of Ω*Ω→R, satisfying the condition:

(1)d(x,y)>=0 , x,y∈Ω

(2)d(x,y)=0If and only if x=y

(3)d(x,y)=d(y,x), x,y∈Ω

(4)d(x,y)<=d(x,z)+d(z,y), x,y,z∈Ω

This is the definition of distance that satisfies positive definiteness, symmetry and triangular inequality. Minkowski distance is the most commonly used measure of distance, i.e.

$$d_q(x,y) = [\sum_{k=1}^{p}|x_k - y_k|^q]^{1/q} \text{ , q>0}$$

When q=1,2 or q→∞, we get:

Absolute distance

$$d_1(x,y) = |x_k - y_k|$$

Euclid distance

$$d_2(x,y) = [\sum_{k=1}^{p}|x_k - y_k|^2]^{1/2}$$

Chebyshev distance

$$d_\infty(x,y) = \max_{1\le k\le p}|x_k - y_k|$$

Euclid distance is the most commonly used in Minkowski distance, and its advantage is that Euclid distance remains unchanged when the coordinate axis rotates orthogonally. The Minkowski distance needs to adopt the same dimension, that is, it needs to standardize the data.

Mahalanobis distance

$$\mathrm{d}(x,y) = \sqrt{(x-y)^T\sum^{-1}(x-y)}$$

Mahalanobis distance is invariant to all linear transformations, so it is not affected by dimensions.

Similarity measure between classes

If there are two sample classes G_1 and G_2, we measure the distance between them by the following method.

Shortest distance method

$$\mathrm{D}(G_1,G_2) = \min_{\substack{x_i\in G_1\\ y_i\in G_2}}\{d(x_i,y_i)\},$$

Maximum distance method

$$\mathrm{D}(G_1,G_2) = \max_{\substack{x_i\in G_1\\ y_i\in G_2}}\{d(x_i,y_i)\},$$

Gravity center method

$$\mathrm{D}(G_1,G_2) = \mathrm{d}(\bar{x},\bar{y}),$$

Class average method

$$\mathrm{D}(G_1,G_2) = \frac{1}{n_1 n_2}\sum_{x_i\in G_1}\sum_{x_i\in G_2} d(x_i,y_i),$$

It is equal to the average distance between two sample points in G_1 and G_2, and n_1 and n_2 are the number of sample points in G_1 and G_2 respectively.

Sum of squares deviation method

$$D_1 = \sum_{x_i \in G_1} (x_i - \bar{x}_1)^T (x_i - \bar{x}_1)\ , D_2 = \sum_{x_j \in G_2} (x_j - \bar{x}_2)^T (x_j - \bar{x}_2),$$

$$D_{12} = \sum_{x_k \in G_1 \cup G_2} (x_k - \bar{x})^T (x_k - \bar{x})$$

$$\mathrm{D}(G_1, G_2) = D_{12} - D_1 - D_2$$

The shortest distance and longest distance method, as the name implies, takes the shortest/long distance between samples in the two categories as the distance of the two categories; The center of gravity method is to take the distance between the two kinds of physical centers of gravity as the distance of the two kinds; Class average method is to take the average value of the distance sum between all samples in two classes as the distance between classes; In the sum of squares of deviation, D_1 and D_2 are the distances between two kinds of internal samples, while D_{12} can describe the distances between two kinds of samples. $\mathrm{D}(G_1, G_2) = D_{12} - D_1 - D_2$ According to the formula, the smaller the distance between two kinds of internal points, the greater the separation distance between two kinds of internal points, the greater the D value. It can be seen that the sum of squares of class difference method is easier to distinguish each kind of cluster.

3.3 R type clustering

Variable similarity measure

In the application of clustering, it is also very important to cluster indicators. This method can help us analyze the correlation degree of each indicator and intuitively see the relationship between each indicator. We can filter the indicators according to the results of clustering, find out the most important influencing factors (indicators), eliminate the indicators with high similarity, and select representative indicators for the next analysis and prediction, thus saving a lot of time and avoiding repeated work. Therefore, the similarity of indicators (variables) is an important indicator to distinguish various types. The higher the similarity, the easier it is to be grouped into one category. The following are several representative similarity measurement methods.

(1) Correlation coefficient

$$r_{jk} = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\left[\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2\right]^{\frac{1}{2}}} \quad ,$$

In clustering analysis of variables, correlation coefficient matrix is used most.

(2) Cosine of included angle

The cosine r_{jk} of the angle between two variables x_j and x_k can be used to define their similarity measure.

$$r_{jk} = \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{\left[\sum_{i=1}^{n} {x_{ij}}^2 \sum_{i=1}^{n} {x_{ik}}^2\right]^{\frac{1}{2}}}$$

Variable clustering

Similar to the class distance calculation method of sample set clustering analysis, the index (variable) clustering can also adopt the longest distance method, the shortest distance method and so on.

Maximum distance method

$$D(G_1, G_2) = \max_{\substack{x_i \in G_1 \\ y_i \in G_2}} \{d_{ik}\}$$

Shortest distance

$$D(G_1, G_2) = \min_{\substack{x_i \in G_1 \\ y_i \in G_2}} \{d_{ik}\}$$

Among them $d_{ik} = 1 - |r_{jk}|$ or $d_{jk}^2 = 1 - r_{jk}^2$.

4. COMMON CLUSTERING ALGORITHMS

4.1 Hierarchical clustering

Here is an introduction to agglomeration-level clustering (AGENES) that adopts a bottom-up strategy. Taking Q-type clustering as an example, we first preprocess the sample to abstract the sample into n points in a k-dimensional space, n points are n different samples, and the index of the sample is k. We select the corresponding distance calculation formula, and first calculate the distance between each point and all other different points. Then, the shortest distance among these distances is selected, and two samples connected by the shortest distance are classified into one class, and the distance between these two samples is taken as the common platform height of these two samples in the cluster diagram. Then select the calculation method of inter-class distance, and recalculate the distance between this class and other samples. After calculation, return to the first step, select the minimum value of all sample class distances, merge and set the platform height, recalculate the distance, and reciprocate until all samples are grouped into one class.

The output is a cluster diagram. We can analyze the cluster diagram and select the number of cluster categories according to the level of various platforms in the cluster diagram. The greater the platform height difference, the greater the gap between the two categories. Selecting a fixed platform height in the cluster diagram can complete the selection of cluster categories, as shown in the figure:

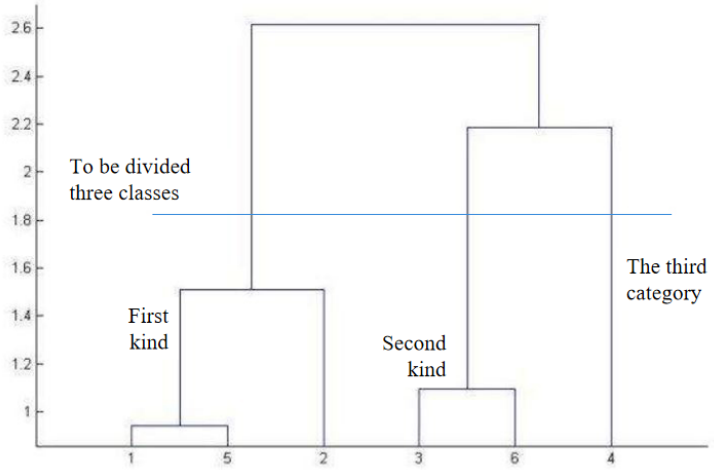

Figure 1 Hierarchical clustering diagram

Strict procedure steps are as follows: set Ω=$\{w_1, w_2, w_3, \ldots, w_n\}$

(1) Calculate the cluster$\{d_{ij}\}$ between every two n sample points, and record it as matrix D=$(d_{ij})_{n*n}$

(2) Firstly, N classes are constructed, each class contains only one sample point, and the platform height of each class is 0.
(3) The nearest two classes are merged into a new class, and the distance between these two classes is taken as the platform height of cluster graph.
(4) Calculate the distance between the new class and the current class. If the number of classes is equal to 1, go to step (5), otherwise, go back to step (3).
(5) Draw a cluster diagram.
(6) Determine the number of classes and classes.
Obviously, the result of this clustering method is related to the distance formula between selected samples and classes and the number of selected classes. The hierarchical clustering method described here is mainly bottom-up clustering. At first, all samples are assumed to belong to one class, then the most similar samples are merged and iterated continuously, and finally all samples are clustered into one class. This method can cluster samples with multiple indicators or multiple indicators (variables) intuitively and simply, and choose the number of clusters independently according to the cluster diagram and sample information after drawing the cluster diagram. It is a simple and easy clustering method.
This clustering method is suitable for dealing with the sample clustering problem with unknown clustering number and multiple indicators. For example, it is required to cluster samples and grade salespersons according to multiple sales performance indicators. Hierarchical Q-type clustering can be used, and indicators should be clustered. By analyzing the data of multiple education level indicators in various regions, the most representative indicators can be selected, and hierarchical R-type clustering can be used. The algorithm complexity of analytic hierarchy process is O(n^3), so it is not suitable for processing large amounts of data. The more data, the longer the algorithm needs. When there are more data, k-means clustering algorithm can be used.

4.2 K-means clustering

K-means clustering is a typical clustering method based on distance. Firstly, sample space is constructed, k sample points are selected, and these k points are used as cluster centers of k classes. The distances between other samples and these k cluster centers are calculated, and other points are divided into the classes where the nearest cluster centers are located. Then, recalculate the center points of k classes, recalculate the distances from all samples to k cluster centers, and subdivide all samples. Iteration is repeated until a certain termination condition is met, iteration is stopped, and clustering is completed. Termination conditions can be the following three:
No sample points were reassigned to different classes
Cluster center no longer changes
The sum of squares of errors is locally minimum
K-means algorithm steps:
Select k samples from n data samples as initial clustering centers.
According to the mean value (center object) of each cluster sample, the distance between each sample and these cluster centers is calculated, and the corresponding samples are subdivided according to the minimum distance.

Recalculate the mean value (center object) of each cluster until the cluster center no longer changes. This division minimizes the following equation:

$$\mathrm{E} = \sum_{j=1}^{k} \sum_{x_i \in w_j} \left|\left| x_i - m_j \right|\right|^2$$

x_i is the position of the ith sample point; m_j is the position of the j th cluster center

Loop steps (2) and (3) until each cluster is no longer changed.

K-means algorithm is a typical clustering algorithm based on distance, which uses distance as the evaluation index of similarity, that is, the closer the distance between two samples, the greater their similarity. The ultimate goal of this algorithm is to get compact and independent categories.

Output cluster diagram is as follows:

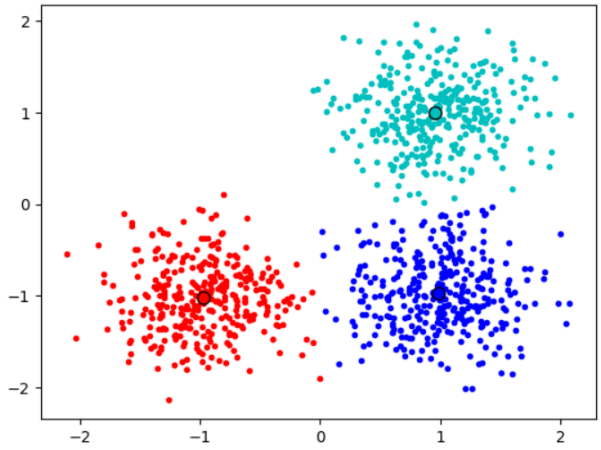

Figure 2 K-means clustering result graph

K-means algorithm needs to set the K value in advance before calculation, that is, the number of categories in clustering results, so special methods are needed to determine the most suitable K value in order to achieve the best effect of k-means algorithm. A simple and fast calculation method is to divide the number of samples by 2 and open the value as k value. For easily observed data (such as two-dimensional data), we can find the number of major categories through observation, so as to find a suitable k value. For data that are not easy to observe, if the number of data is limited, we can set the sum of error squares, try several different k values, and finally select the k value with the smallest error as the appropriate k value. If the amount of data is large or the requirements are strict, we can use Elbow Method to determine the value of k.

The elbow rule refers to the calculation error and the formula f(x)=SSE(x). There is a certain relationship between k and SSE. When x<k, adding 1 to x will cause the SSE value to change greatly. When x>k , x plus 1 results in a small change in the cost value, and the correct k value is at that turning point. This method is suitable for the case of small k value, and it is not always applicable, because the cost curve is named after the elbow.

Error leveling method and calculation formula of SSE(sum of the squared erroes) are as follows:

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{p \in c_j} |p - m_i|^2$$

p is the sample point position, m_i is the cluster center position of each class, and c_j is each class.

The time complexity of k-means algorithm is o(IKN), where n is the number of sample points, k is the number of central points and I is the number of iterations. k-means algorithm has the advantages of fast and simple operation, good clustering effect, and is suitable for high dimensions, while its disadvantages are sensitive to outliers, noise points and isolated points, the number of clusters k needs to be set in advance, and the selection of initial clustering centers and different initial points may lead to completely different clustering results.

4.3 Mean shift clustering

Mean shift clustering is an average algorithm based on sliding window, which constantly seeks the area with the highest density in sample data points. The two-dimensional representation of this method is vivid. Firstly, data points are placed in a two-dimensional plane, a certain number of points are randomly selected in the space, and a circular window is drawn with these points as the center and a fixed value as the radius. The points with the highest density of data points in the window are calculated, and the previous points are replaced with the points with the highest density. Then, the circular window is drawn with the fixed value as the radius, and iteration is repeated. In this way, the window will move to the location with dense data points until it reaches the location with the densest data points. When the windows overlap or all the windows are no longer moving, the algorithm will finish clustering, and the center point of the window is the center point of clustering. Obviously, this algorithm uses the idea of hill climbing, which makes the center of the window move to the place with high data density, and finally achieves the goal of clustering.

The specific steps are as follows:

(1)Firstly, a circular sliding window is made with a randomly selected point as the center r as the radius. Its goal is to find the highest density point in data points and take it as the center;

(2)After each iteration, the center of the sliding window will move in the direction of thinking about higher density;

(3)Continue to move until the number of middle points in the sliding window can not be increased by moving in any direction, and then the sliding window converges;

(4)The above steps are carried out on multiple sliding windows to cover all points. When the last sliding window converges and overlaps, the points it passes through will be clustered into a class through its sliding window;

Compared with k-means algorithm, this method does not need to specify the number of clusters in advance, and the number of clusters is determined by itself, which is relatively in line with the results of intuitive cognition. The disadvantage is that the selection of sliding window radius has a great influence on clustering results. When the sample contains multiple attributes, this method should be used to preprocess the index value appropriately to adapt to the selected window radius.

5. APPLICATION OF CLUSTERING

In the commercial field, cluster analysis can divide the target groups into groups with multiple indicators, and identify the categories with typical characteristics through classification, so as to formulate corresponding business strategies, carry out personalized and refined operations, services and product support; In the data analysis, we can also find abnormal points and abnormal data through clustering, and carry out targeted processing; In social investigation, we can analyze a large number

of collected data through clustering, so as to distinguish good from bad and establish an evaluation system; For a large number of data with multiple indicators, we can select the most representative indicators for analysis through clustering before investigation and analysis, saving a lot of energy and resources; There are many applications of clustering. Nowadays, we are entering the network information age, and big data explosion is common. Cluster analysis is an important tool for big data analysis, data mining and machine learning.

6. SUMMARY

The first part of this paper mainly discusses the basic concepts of clustering algorithm, the difference with classification, the basic process and evaluation criteria. Only by understanding the concept of clustering and the difference between clustering and classification can we understand the purpose and principle of clustering more clearly. In the second section of this paper, the classification of clustering is discussed, which is mainly divided into Q-type clustering and R-type clustering, and several formulas for calculating the distance of sample points in clustering are introduced. Being familiar with the classification and distance calculation formulas of clustering will help us to make good use of clustering, and selecting suitable algorithm categories and distance formulas by analyzing application scenarios will help us better cluster analysis. At the end of this paper, three commonly used clustering algorithms are introduced, which are hierarchical method, k-means clustering and mean shift clustering. Among them, k-means clustering is widely used, and we also provide a variety of reference methods for the selection of K value. In a word, this paper introduces the basic concept and classification of clustering algorithm, and the principles and steps of some commonly used clustering algorithms such as k-means algorithm.

REFERENCES

[1]. Li Lingling, Research on Time Complexity of Condensed Hierarchical Clustering, Journal of Suzhou University. Vol. 26, No. 2, 2011: 21-22
[2]. Si Shoukui, Sun Zhaoliang, Mathematical Modeling Algorithms and Applications (Second Edition)
[3]. MATLAB Mathematical Modeling Method and Practice (Third Edition)
[4]. Five clustering algorithms that must be familiar in data science: https:// baijiahao.baidu.com/s?id =1625408992304959354&wfr=spider&for=pc
[5]. The concept, process and algorithm of clustering:http://blog.sina.com.cn/s/blog_78bdf8e801013 wx6.html
[6]. Detailed explanation of machine learning k-means algorithm:https://blog.csdn.net/sinat_303532 59/article/details/80887779?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-10&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-10
[7]. The choice of k value of k-means algorithm:https://www.biaodianfu.com/k-means-choose-k.html
[8]. Typical applications of cluster analysis: https://www.jianshu.com/p/7b13610eb674