Advances in Engineering Research (AER), volume 82
2016 International Conference on Engineering and Advanced Technology (ICEAT-16)

Research on clustering algorithm

Rui Wang [1,a], Jinguo Wang [2,b*], Na Wang [3,c]
[1]Department of Information Engineering,Jilin Business and Technology College China
[2]Department of Urology, the First Hospital of Jilin University,China
[3]Department of Anaesthesiology, the First Hospital of Jilin University, China
[a]Xiaoben6666@126.com,[b]wangjinguolily@163.com,[c]lilyly12345@163.com
*corresponding author

Key words: Clustering algorithm, Minkowski distance, Data mining

Abstract. Clustering is one of the important techniques of data mining. It can be divided data into several classes or clusters according to certain rules, which makes the data objects of the same class have high similarity, and the different data objects are very different. In this paper, the clustering algorithm is analyzed and compared in detail, and summarize advantages and disadvantages of this algorithm.

1. Introduction

The clustering algorithm inspired by the foraging process of ants is also called the clustering algorithm based on the principle of ant foraging[1][2]. The ant foraging behavior is divided into two parts: the search for food and the handling of food[3]. At the same time, the data object is regarded as an ant, and the clustering center is regarded as the "food source". The clustering process of the data object can be transformed into the process of the ants' foraging. Under the guidance of pheromone, the ants can complete the clustering of data objects[4]. A new ant colony clustering algorithm is proposed in this paper.

2. Clustering algorithm

Clustering analysis is an important field of data mining, and it is also one of the hot issues in the current research[5]. It is the technology which studies the logical or physical relations between data, and divides the data set into several classes, which are composed of similar data points in nature. The result of cluster analysis can not only reveal the inherent relation and the difference between the data, but also provide an important basis for further data analysis and knowledge discovery, such as association rules between data classification and data trend.
The ant colony algorithm is a swarm intelligence optimization algorithm has good global search ability in solving solutions, many complicated optimization problems have shown excellent performance and great potential for development, has become a concern of the frontier[6][7]. The ant colony algorithm is applied in the field of cluster analysis, massive data processing, is a new kind of intelligent information age knowledge discovery, in the analysis and comparison of other classical clustering algorithm, gradually shows its vitality, highlighting its unique advantages.

3. Classification of cluster analysis

According to the clustering criteria, the clustering method can be divided into the following

two types[8][9].

Statistical clustering method. Statistical clustering based on geometric distance between objects. Statistical clustering analysis includes the system clustering method, the decomposition method, the addition method, the dynamic clustering method, the ordered sample clustering method, the overlapping clustering and the fuzzy clustering and so on. This clustering method is based on the overall comparison of the cluster, it needs to study the classification of all the individuals in order to determine the classification[10].

Concept clustering method. The concept clustering method based on the concept of object has the concept of clustering. The distance here is no longer a geometric distance in the traditional method, but is determined according to the description of the concept[11].

4. Distance and similarity measure

The quality of a clustering process depends on the choice of the metric, and therefore must be carefully chosen to measure the standard. Here are a brief introduction to these standards.

distance function. According to the distance axiom, four conditions are required to satisfy the distance axiom in the definition of measure: the self similarity, the minimum, the symmetry and the triangle inequality. Commonly used distance functions are as follows:

(1)Minkowski distance[12]

Assumed x,y is the corresponding feature, n is the dimension of the feature. The Minkowski distance of x, y is as following:

$$d(x,y)=[\sum_{i=1}^{n}|x_i-y_i|^r]^{1/r} \qquad (1)$$

If R =1, Minkowski absolute distance is

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i| \qquad (2)$$

If R =2, Minkowski Euclidean distance is

$$d(x,y)=[\sum_{i=1}^{n}|x_i-y_i|^2]^{1/2} \qquad (3)$$

(2)Quadratic distance[13]

The form of the two distance measure is as follows:

$$d(x,y)=((x,y)^T A(x-y)) \qquad (4)$$

In which, A is the non negative definite matrix. When the A is the unit matrix, the Quadratic distance is the Euclidean distance. When the A is a diagonal matrix, the Quadratic distance is weighted Euclidean distance.

$$d(x,y)=[\sum_{i=1}^{n}a_{ij}|x_i-y_i|^2]^{1/2} \qquad (5)$$

5. Partition clustering method

Main idea. Given a data set of n, partition clustering technology will construct data M partition,

each partition represents a cluster, and m<=n. The data is divided into M clusters, and it must satisfy: Each cluster contains at least one object. Each object must belong to only one cluster.
Given the number of partitions M, the partitioning method first creates an initial partition. Then an iterative relocation technique is used to improve the classification by using the object in the division of the partition. A good division of the general standard is: the same class of objects between the best possible "similar or related", but not the same kind of objects in the asked as far as possible, "away" or different[14].
Evaluation function. Most of the evaluation functions for clustering design are mainly two aspects: each cluster should be compact, and the distance between each cluster should be as far as possible. A direct method for the realization of this concept is to observe the within cluster variation and between cluster variation. Within cluster variation($w(C)$) measures the compactness of clustering, between cluster variation($b(C)$) measures the distance between different clusters.
Within the class difference can be defined using a variety of distance functions, the simplest is to calculate the class of each point to its class center of the square and the distance[15]:

$$w(C) = \sum_{i=1}^{k} w(C_i) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, \overline{x_i})^2 \quad (6)$$

The difference between cluster variation is defined as the distance between cluster centers:

$$b(C) = \sum_{1 \le j \le i \le k} d(\overline{x_i}, \overline{x_j})^2 \quad (7)$$

6. Comparison of clustering algorithms

With all kinds of clustering algorithms have been proposed, each kind of new algorithms are claimed to be superior to the previous one, which makes the comparison between the various clustering algorithms become more and more difficult[16]. In fact, no algorithm can be proved to be superior to all other algorithms in every aspect. Moreover, many algorithms are designed for a specific domain[17].
There are six criteria to measure a clustering algorithm. Ability to handle large data volumes. If we can find clusters of arbitrary shapes. Whether it’s sensitive to a large amount of noise in the data. Whether it’s sensitive to the data input sequence. Whether have the ability to handle high dimensional data. Number of input parameters.

7. Summary

In many practical applications, the similarity between the data objects in the same cluster is very high. They have the same or similar properties in some respects. So they can be used as a whole to deal with or analysis, which is also the practical significance of clustering activities. By clustering the data objects, we can find the relationship between the distribution pattern of data objects and the value of the data attributes.
In this paper, the cluster analysis method is studied and analyzed deeply. The data representation, similarity measure method and main clustering algorithm in cluster analysis are

introduced in detail. The advantages and disadvantages of the main clustering algorithms are analyzed.

References

[1] McInerney T,Terzopoulos D.Topologically adaptable snakes. Proceedings of the Fifth International Conference on Computer Vision(ICCV' 95) . 1995
[2] R. Van Driessche,D. Roose.An improved spectral bisection algorithm and itsapplication to dynamic load balancing. Parallel Computing . 1995
[3] Likas A.The global k-means clustering algorithm. Pattern Recognition . 2003
[4] Fowlkes C,Belongie S,Malic J.Efficient spatiotemporal grouping using the NystrOm method. IEEE Transactions on Pattern Analysis and Machine Intelligence . 2004
[5] Abutaleb A S.Automatic thresholding of gray-level picture using two-dimensional entropies. Pattern Recognition . 1989
[6] Andrew Mehnert,Paul Jackway.An improved seeded region growing algorithm. Pattern Recognition . 1997
[7] Comaniciu Dorin,Meer Peter.Robust analysis of feature spaces: color image segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition . 1997
[8] Bradley P S,Fayyad U M.Refining Initial Points for K-Means Clustering. Proceedings of the 15th International Conference on Machine Learning (ICML'98) . 1998
[9] GIVONI I E,FREY B J.Semi-supervised affinity propa-gation with instance-level constraints. Proceedingsof12th International Conference on Artificial Intelligenceand Statistics(AISTATS) . 2009
[10] Delbert Dueck.Affinity Propagation:Clustering Data by Passing Messages. 2009
[11] Ulrike Luxburg. A tutorial on spectral clustering[J]. Statistics and Computing . 2007 (4)
[12] Zheng Tian,XiaoBin Li,YanWei Ju. Spectral clustering based on matrix perturbation theory[J]. Science in China Series F: Information Sciences . 2007 (1)
[13] An Experimental Comparison of Model-Based Clustering Methods[J]. Machine Learning . 2001 (1)
[14] Tian Zhang,Raghu Ramakrishnan,Miron Livny. BIRCH: A New Data Clustering Algorithm and Its Applications[J]. Data Mining and Knowledge Discovery . 1997 (2)
[15] Remi Ronfard. Region-based strategies for active contour models[J]. International Journal of Computer Vision . 1994 (2)
[16] Michael Kass,Andrew Witkin,Demetri Terzopoulos. Snakes: Active contour models[J]. International Journal of Computer Vision . 1988 (4)
[17] Stephen C. Johnson. Hierarchical clustering schemes[J]. Psychometrika . 1967 (3)