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

Research on Improvement of k-means Clustering Algorithm

Jian Di [a], Kai Yang [b]

School of Control and Computer Engineering, North China Electric Power University, Baoding, China;

[a]dijian6880@163.com, [b]yangkai920930@126.com

Abstract: nowadays, with the rapid development of the information age, make the development of social industry fields more dependent on information data, at the same time, it is affected by factors such as the variety of information and the variety of forms, there are many difficulties and challenges in implementing effective information data acquisition. Under this influence, the clustering algorithm is gradually introduced into the information data mining work due to its unique operational characteristics, which has become one of the reliable methods for obtaining useful information data, this also makes the cluster form including k-means clustering algorithm fully promoted and applied in various economic industry fields, and provides reliable and accurate information data reference for market economy construction and development plays an important role.
Keywords: Clustering; k-means clustering algorithm; improvement measures.

1. INTRODUCTION

Due to the rapid development of electronic information technology and internet technology, it has brought important advantages to the economic construction progress of various industries while deepening the application of technology, providing reliability and accuracy for long-term planning in the industry. The information data has strengthened the efficiency of information circulation between regions, and promoted the information data mining from the scientific and technological research and development to the socialization of the society. At this stage, with the development of social construction, it has entered a new era. The total amount of information data has shown an increasing trend. How can we make full use of clustering ideas to make use of the advantages of k-means clustering algorithm based on existing information resources and providing a reliable data reference for the rational classification and optimal distribution of information has become one of the main tasks of current information development.

2. THE MAIN CONTENT AND CHARACTERISTICS OF CLUSTERING IDEA

The clustering idea is a unique form of data acquisition formed during the reading of information data,the main content is a process of dividing a collection of physical or abstract objects into multiple classes consisting of similar objects. The cluster generated by the cluster is a collection of data objects that are related to each other in the same cluster. Similar to the objects in other clusters, the ideology

formed by the classification and aggregation of various types of information data in the process of industry development based on the above process is called clustering thought. The main features are as follows. Aspects: On the one hand, information data classification is scalable, and different sizes of data information can be selected as basic operational data for different data classification sets; on the one hand, the ability to process different types of data, due to the wide range of applications applied by clustering In the current social production operations, clustering is used to enable it to have comprehensive capabilities for processing data structures in different industries. On the other hand, high-latitude data processing, a database or data warehouse may contain several dimensions or attributes, many clustering algorithms are good at processing low-dimensional data, which may involve only two to three dimensions, and new algorithms including k-means clustering can break the limits of high-latitude data operations, and make it can face more difficult computing challenges in the data operation process [1].

3. MAIN CONTENTS AND CHARACTERISTICS OF K-MEANS CLUSTERING ALGORITHM

K-means algorithm is a classic clustering algorithm based on splitting method in data mining technology, the algorithm is widely used Because of its reliable theory, simple algorithm and rapid convergence. K-means algorithm is a segmentation clustering algorithm based on Euclidean distance, the basic ideas are mainly as follows: C divisions of data based on the number of clusters C, calculate the class core of each division, update the catagory of data to the current division, constantly iteratively adjust clusters and their class cores, until the generics of all data no longer change. From the perspective of algorithm steps, it is mainly divided into the following four steps: In the first step, the number C of clusters is determined within the scope of the data analysis, and C data is randomly selected as the center V1 of the cluster, and the necessary data parameters required for the algorithm are initialized; The second step is to update the cluster, calculate the distance of all the data to the C centers vi, select the nearest class core for each data, and classify the data into the class; In the third step, the cluster center is updated, and the eigenvalues of the same type of data are averaged to obtain an updated cluster center according to the data type attribute represented by each data; The fourth step, iteration: calculating the value of the corresponding objective function of the division, repeating the steps between the second step and the fourth step until the value of J does not change or the J change value reaches a specified smaller threshold [1].

For example: suppose we extract the collection of raw data to (x_1, x_2, ..., x_n),

and each xi is a vector of d dimensions, the purpose of K-means clustering is to divide the raw data into k classes S = {S_1, S_2, ..., S_k} given the number of classification groups k($k \leq n$), on the numerical model, that is, the minimum value of the following expression is as shown in the following figure:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Here μ_i represents the average value of the classified S_i.

However, the dependence of the clustering results of the K-means algorithm on the initial clustering center leads to unstable clustering results, moreover, ignoring the influence of different attributes of data objects on the degree of difference between objects based on the Euclidean distance between

objects also makes the clustering accuracy of K-means algorithm lower. In practical applications, if the algorithm can be improved from both the selection of the initial cluster center and the revision of the Euclidean distance calculation formula between objects, it will be of great significance to reduce the volatility of the traditional k-means algorithm clustering results and obtain a better clustering effect [2].

4. MAIN MEASURES TO IMPROVE K-MEANS CLUSTERING ALGORITHM

In order to further ensure the accuracy and reliability of the K-means clustering algorithm, it should first be determined that the density parameter of each data object is determined within the information data cluster, data collection points with small distribution range are used as the initial clustering center of such clustering algorithm, thereby further improving the accuracy and stability of data information. Usually, the greedy algorithm can be used to classify and divide the selected data samples, and form a plurality of data sets, and select an average value as the initial cluster center, at the same time, through the above understanding of the basic content of the K-means clustering algorithm, the data structure of the minimum spanning tree is used to successively add two data sets at the farthest distance to each data set far away from the cluster center, iteration like this until the cluster contains multiple cluster cores, and achieves good results.

Secondly, the original information data is initialized, and the processed information data is regarded as a data sample in the K-means clustering algorithm, which promotes the problem of low computational efficiency caused by the reduction of information between samples, based on factor analysis, data mining under complex parameter variables effectively reduces redundant fields and improves the efficiency of K-means grouping algorithm. Literature [8] uses information entropy to weight the attributes of data objects, and uses weights to modify the distance calculation formula, which improves the accuracy and stability of K-means clustering to some extent [4].

Thirdly, in the process of continuous clustering practice, a K-means clustering algorithm improved by entropy method and dynamic programming algorithm is proposed. The algorithm uses the entropy method to determine the weight of the data attribute and further obtain the weight coefficient between the adjacent data sets, and gradually use the weighted Euclidean distance as the reference standard for measuring the similarity between data sets, and use the dynamic programming algorithm to obtain the distance accumulated and the largest K data objects as the initial clustering center. The application results of the algorithm in mine monitoring sensor clustering show that the algorithm improves the accuracy and stability of clustering [5].

5. SUMMARY

In summary, the K-means clustering algorithm has an important influence on improving the efficiency and accuracy of information data acquisition. Through the K-means clustering algorithm, the required data clusters are classified and operated in a dynamic manner. The clustering center of the information data set is determined, and the distance calculation formula is modified by the improved weight to improve the accuracy of the cluster, and the K-means clustering algorithm is improved to adapt to the complex and variable information data environment.

REFERENCES

[1] Li Hui, Shi Zhao, Yi Junkai, et al. Secondary clustering recommendation algorithm based on information entropy[J]. Computer Engineering, 2016,42(5):213-217,223. (in Chinese)

[2] He Miao. Concept lattice compression based on clustering thought [J]. Journal of Shaanxi University of

Technology (Natural Science Edition), 2016,32(3):78-82. (in Chinese)
[3] Fan Huitao, Feng Tao. The Weighted Conditional Entropy and Attribute Reduction Based on Clustering[J]. Journal of Zhengzhou University, 2018,50(1):39-46. (in Chinese)
[4] Wang Qian, Wang Cheng, Feng Zhenyuan, et al. A Survey of K-means Clustering Algorithms[J]. Electronic Design Engineering,2012,20(7):21-24. (in Chinese)
[5] Lu Ruiqiang, Ma Fumin, Zhang Tengfei. Rough K-means clustering algorithm based on interval 2-type fuzzy metric[J]. Rough K-means clustering algorithm based on interval 2-type fuzzy metric, 2018,31(3):265-274. (in Chinese)