Computer and Information Science; Vol. 12, No. 1; 2019
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Research and Improvement Method Based on k-mean Clustering Algorithm

Guohua Zhang[1], Kangting Zhao[1] & YiLi[1]

[1] Hunan University of Technology, China

Correspondence: Guohua Zhang, Hunan University of Technology, Zhuzhou, Hunan, 412000, China. Tel: 86-139-0733-5354. E-mail: 9884188@qq.com

Received: July 11, 2018 Accepted: August 16, 2018 Online Published: January 18, 2019

doi:10.5539/cis. v12n1p49 URL: https://doi.org/10.5539/cis.v12n1p49

Natural Science Foundation Project of Hunan Province(2017JJ2070);Research-based Learning and Innovative Pilot Program for College Students in Hunan Province in 2018（XJT-2018-255）

Abstract

In view of the sensitivity of the traditional mean algorithm to outliers and noise points, an improved mean algorithm is proposed in this paper, which is based on the density of the distribution of objects in space. In the measurement of density, the sensitivity of clustering effect to initial parameters is reduced. The improved algorithm can filter the "noise" data and discover the clustering of arbitrary shapes, which is obviously superior to the standard mean algorithm.

Keywords: Clustering Analysis, K-mean Algorithm, OPTICS Algorithm

1. Introduction

With the great progress of science and technology, people in real life and work will frequently face the embarrassment of having a lot of complicated data information and can not effectively and accurately extract valuable information, data mining theory and technology came into being. Data mining is the process of mining interesting patterns and knowledge from a large amount of data, obtaining useful information, implicit connections and potential laws to guide people to discover interesting and regular data in a large number of data, and to help researchers make decision analysis. Clustering analysis is an important technical means of data mining, it is a process of dividing dataset objects or observations into several subsets, and it is an unsupervised learning method (Sun & Liu, 2008). Clustering analysis requires the similarity of objects in the class divided by clustering, and the difference of objects between classes is large. The k-mean clustering is one of the classical clustering algorithms, and the Euclidean distance is often divided as the criterion of similarity, the lower the distance and the smaller the class, the more compact the generated clustering results and the more independent the class, the better the clustering effect (Han, Kamber & Pei, 2012). The k-mean clustering algorithm is easy to divide, and the convergence speed is fast. However, k-mean clustering does not necessarily converge on the global optimal solution, often converges to local optimization, clustering results often appear unstable phenomenon, affecting clustering effect, thus affecting people to obtain effective information in the data and decision making formulation (Kang, Sand & Yip, 2017). Many researchers use evolutionary algorithms and group intelligence methods and k-mean mixed clustering, as well as from the perspective of Initialization Clustering center and optimization of Clustering center position, and improve the k-mean clustering.

2. The k-mean Clustering Algorithm

2.1 Related conception

Clustering analysis, as one of the main functions of data mining, is mainly divided into unclassified sample types, which can effectively promote the collation and classification of samples. Clustering analysis is to use the principle of things as the basic guiding ideology method to divide the data, the data according to a certain rules into different groups or classes, can also be several clusters, and make relatively close to the sample or similar samples in the same cluster, the non-similar samples are classified into another different cluster. This allows a group of physical or abstract objects to be divided into groups based on the degree of similarity between them (distance), in which similar objects form a group, a process known as a clustering process. There are many kinds of clustering

algorithms, and the k-mean clustering algorithm based on partitioning method is a simple and widely used method in clustering algorithm.

Given the number to be divided, the method based on partitioning first creates an initial division, and then uses an iterative repositioning method to try to improve the quality of the division by moving objects between divisions.

The k-mean basic idea of the algorithm is that first randomly select a point as the initial clustering center, the distance of each object to all clustering centers is then calculated, and the data object is grouped into the class where the clustering center closest to it is located. The similarity of the samples in the mean is determined by the distance between them, and the closer the distance, the higher the similarity, the lower the similarity is indicated. The value of similarity is usually represented by the inverse of the distance. Among them, the common distance calculation method has the European distance and the Manhattan distance. In cluster analysis, two m dimension samples are set up $\mathrm{x}_i = (x_{i1}, x_{i2}, \cdots, x_{\mathrm{im}})$ and $\mathrm{x}_{\mathrm{j}} = (x_{j1}, x_{j2}, \cdots, x_{jm})$, then there is the following relational formula:

European distance is the calculation of the distance between two points in the European space, is the most easy to understand the distance calculation method, which is calculated as follows:

$$\mathrm{d}ist_{ed} = \sqrt{\sum_{k=1}^{\mathrm{m}} (x_{ik} - x_{jk})^2} \tag{1}$$

Manhattan distance is also known as the city block distance, the European distance indicates the straight distance between two points in the space, but the actual distance between the two locations in the city is the distance to travel along the road, rather than calculating the straight distance directly through the building, Manhattan distance is used to measure such actual travel distance.

$$\mathrm{d}ist_{mand} = \sum_{k=1}^{\mathrm{m}} \left| x_{ik} - x_{jk} \right| \tag{2}$$

Among them, the European distance is more commonly used.

2.2 Steps of Algorithm

The k-mean clustering algorithm to divide the X dataset into k clusters that $C = \{C_1, C_2, \cdots, C_{\mathrm{k}}\}$. The center of each cluster is $c_j (j = 1,2,\cdots,k)$. K cluster must meet these conditions: C_j non-empty, $\bigcup_{\mathrm{j}=1}^{k} C_j = X, i \neq j, C_{\mathrm{i}} \cap C_j = \phi.$ According to such partitioning conditions, in the clustering results, the objects of the same clustering are as similar as possible, while the objects in different clustering ar e as large as possible, so as to make it more recognizable and differentiated.

Here, the label for each cluster is not given in advance, so this is a problem of unsupervised learning. x_i in the X dataset are represented $\left\| x_i - \mathrm{c}_j \right\|$ by the European distance c_j from the cluster center. Theo perating steps of the algorithm are described as follows:

(a)Randomly initialize a cluster center point: $c_1, c_2, \cdots, c_k \in R^d$;

(b)Repeat the loop to do the following until the algorithm converges.

For each object x_i in the data:

$$C_j := \arg\min_j \left\| x_i - c_j \right\|^2 \tag{3}$$

For each cluster Center c_j:

$$c_j = \frac{\sum_{i=1}^{m} 1\{C_i = j\}\ x_i}{\sum_{i=1}^{m} 1\{C_i = j\}} \tag{4}$$

In the above algorithm model, step 1 indicates that the data X set randomly selects k object as the initial clustering center. Formulas (3) Divide each data object x_i into the category of the cluster center c_j closest to it; the formula (4) Moves the cluster center c_j to the mean of all points at the current cluster.

The k-mean algorithm is convergent to a certain extent, and the clustering formula is deformed as follows:

$$J(C_j, c_i) = \sum_{i=1}^{m} \left\| x_i - c_j \right\|^2 \tag{5}$$

$J(C_i, c_j)$ describes the European distance squared sum of the data objects in the current clustering division to the corresponding clustering center. The k-mean clustering effect is evaluated by measuring standard function, and the standard function is as:

$$J_e = \sum_{i=1}^{k} \sum_{x_i \in C_j} \left| x_i - c_j \right|^2 \tag{6}$$

Transform the upper formula J_e to get:

$$\frac{\partial J_e}{\partial c_j} = \frac{\partial}{\partial c_j} \sum_{i=1}^{k} \sum_{x_i \in C_j} \left| x_i - c_j \right|^2 = \sum_{i=1}^{k} \sum_{x_i \in C_j} \frac{\partial}{\partial c_j} \left| x_i - c_j \right|^2 = (-2) \sum_{x_i \in C_j} \left| x_i - c_j \right| \tag{7}$$

If $(-2)\sum_{x_i \in c_j} \left| x_i - c_j \right| = 0$, then $c_i = \frac{1}{|C_i|} \sum_{x_i \in c_j} x_i$. That is, the optimal result is to calculate the mean of t

he cluster. The smaller J_e the result is that the more compact and independent the clustering results, the better the clustering effect.

K-Mean algorithm is a kind of clustering algorithm based on division, which often uses error squared and criterion function as criterion function. Valid if the various types are obvious and J_e data dense. However, if the shape size of each type varies greatly, it will cause the algorithm to converge locally. Therefore, the clustering center may fall into the local optimal solution, and it is difficult to obtain the global optimal solution. Aiming at this problem, a density based OPTICS algorithm is proposed in this paper.

3. Improvement of Density-Based k-mean Clustering Algorithm

3.1 The Fundamental Rules of Arithmetic

Because the k-mean algorithm is sensitive to "noise" points, this section proposes an improved k-mean clustering algorithm, which can filter "noise" data, discover clustering of arbitrary shapes, and provide a method to determine the optimal number of clustering, and find the best clustering center. First, the algorithm selects a point at the farthest distance from the cluster center in a high-density data point, and sees it as a new clustering center, which is placed in a collection of clustering centers. For a dataset, when the optimal clustering number is determined, the clustering center obtained according to the improved algorithm is also determined, so that the stability of the algorithm will be greatly improved.

3.2 Concepts related to OPTICS Algorithms

(1) ε − Neighborhood: For any sample, the ε − neighborhood is a subset of the sample set D and x_j the distance is not greater than ε, that is

$$N_\varepsilon(x_j) = \{x_i \in D \mid \rho(x_i, x_j) \le \varepsilon\} \tag{8}$$

(2) Core object: For any sample x_j, if its ε −neighborhood contains at least M samples, That is $\left|N_\varepsilon(x_j)\right| \geq M$, x_j is the core object.

(3) Core-distance

Assuming that point p contains the minimum radius distance of a M neighbor object is $M - \rho$（p）, t he core distance of p is defined as:

$$\text{core} - \rho(\text{p}) = \begin{cases} Undefined, & \text{p is not core point} \\ \text{M} - \rho(\text{p}), & \text{p is a core point} \end{cases}$$

The core distance is the minimum neighborhood radius at which a point becomes the core point.

(4) Reachability-distance

Assuming that q is a point in the ε −neighborhood of point p, the accessibility distance of q for p is defined as:

$$\text{Reachability} - \rho(\text{q, p}) = \begin{cases} Undefined, & \text{p is not core point} \\ Max(core - \rho(p, q)), & p \text{ is core point} \end{cases}$$

It can be seen that q the reach distance of p is the minimum distance from p direct density up to q. The distance is directly related to the spatial density, and if the point is in a large space density, it can reach a distance from the density of the adjacent point. When the distance value is minimum, it is shown that the clustering effect of the algorithm is the best, when the optimal clustering number is K.

4. Conclusion

This paper mainly improves the traditional k-mean algorithm and improves the clustering effect of the algorithm. The algorithm determines the optimal clustering number according to the distance, gets rid of the control of ε value, and makes the area before and after the expansion closer.

References

Han, J. W., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques [M]. Beijing: Machinery Industry Press.

Kang, S. H., Sand, B. B., & Yip, A. M. (2017). A regularized k-means and multiphase scale segmentation[J]. *Inverse Problems & Imaging, 5*(2), 407-429.

Li, C. S., & Wang, Y. N. (2010). A new method for the initialization of Clustering Center [J]. *Control Theory and Application, 27*(10), 1435-1440.

Li, H. M. (2018). Review of Big Data clustering algorithms [J]. *Wireless Interconnection Technology, 2018*(18), 157-158.

Liao, W. D., Zhu, F. B., Wang, H. Q., & Sun, X. K. (2018). Mean clustering algorithm based on artificial swarm optimization [M]. *Computer Measurement and Control, 26*(4).

Sun, G., & Liu, J. (2008). Research on clustering algorithm [J]. *Software Journal, 19*(1), 48-61.

Copyrights