OPAM Algorithm Based on Density Peak Optimization Initial Center

Guoqing Qiu [a], Wanying Zhao [b] and Shaoyun Zhang [c]

School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.

[a]460557728@qq.com, [b]zhaowying630@163.com, [c]1278762367@qq.com

Abstract

Aiming at the problems of traditional PAM clustering algorithm and OPAM clustering algorithm based on reverse learning, the initial clustering center is sensitive, easy to fall into the local optimum, and the initial value has a great influence on the final center. Combined with the density peak clustering algorithm, OPAM algorithm based on density peak optimization initial center, namely DP-OPAM. The algorithm uses the local density of data points and the shortest distance between these points to a higher density point, and selects the category to which the more dense and nearest data points belong using the decision graph as the initial cluster center. According to the initial clustering center, OPAM clustering algorithm with reverse learning is used to get the clustering result[1]. Comparing the new algorithm with the original OPAM algorithm, the new algorithm can not only automatically determine the clustering center[2], but also improve the accuracy and clustering time.

Keywords

OPAM, Clustering, Density peak,Clustering center.

1. Introduction

Clustering is a process of dividing a data set. In this process, elements in a data set are divided into several unrelated groups or classes that satisfy the condition that the elements of the same group or class , While the elements in different groups or classes are dissimilar[3]. Clustering analysis of the process of division is not artificial, but by data mining clustering algorithm to be divided. After the clustering is completed, the judgment of the quality or the effect of the clustering is also judged by a specific way and method[4]. Clustering can be used as a data mining method in addition to independent data mining tasks, but also can be used as a preprocessing step for other data mining methods, such as characterization, classification and attribute subset selection.

There is also a commonly used analysis method in data mining methods - classification analysis. Classification and clustering analysis are the same in the purpose and result of the algorithm, and they all divide a collection of different elements or data into several subsets. The elements or data in each subsection are similar, There are differences with the elements or data in other subsets[5]. However, it is worth emphasizing that clustering analysis is not only used as a data mining method different from classification analysis, but its special feature is that clustering analysis is an unsupervised learning data mining method. The so-called unsupervised learning method refers to the absence of any reference standard under the premise of the algorithm based on elements or the existence of the data itself is free to divide, do not need to be based on the known reference standard to divide the elements or data, Is a process of automatically discovering packets.

2. PAM Clustering Algorithm

2.1 The idea of PAM Algorithm

The PAM (Partitioning Around Medoid) algorithm is a method of dividing the center point, and is one of the earliest k-medoids algorithms[6]. Is the advantage of the PAM algorithm: the "noise" and outlier data PAM algorithm (from other data points of long-term data) is not sensitive, more robust algorithm; the input sequence and cluster detection data formed by PAM clustering algorithm is

independent; to handle different types of data. The PAM algorithm first selects representative objects in the training data (i.e., data center) and the rest of the object according to the center distance were added to the nearest point center on behalf of the cluster, and then repeated attempts to replace the current center with non center point, calculate the total cost of every replacement, if the total cost negative, non Center Replacement Center, to the rest of the objects according to the distance distribution, if the total cost is in a non center point, calculate the total cost, until the data is no longer change.

2.2 Improved OPAM algorithm

Aiming at the problem that PAM algorithm is easy to fall into local optimum, OPAM algorithm is proposed subsequently, and reverse learning is added based on the original PAM algorithm[7]. The main idea is to set the number of clusters and cluster the training data according to the typical PAM algorithm. In view of the disadvantages that the PAM algorithm is apt to fall into the local optimal solution, the reverse learning algorithm is used to reversely learn each clustering result Inverse clusters, the clustering quality of the original cluster and the reverse cluster are respectively calculated, and the clustering combination with higher quality of clustering is taken.

Assuming a given training data set D needs to be clustered into k clusters, the steps of OPAM algorithm are as follows:

(1) Arbitrarily select an element k from a given data set D, mark the selected element k as the initial representative object or seed o_j;

(2) According to the calculation method of Euclidean distance, calculate the distance between any non-representative object o_i and k representative objects in the data set D and assign o_i to the cluster represented by the proximate object;

(3) Arbitrarily choose a non-representative object o_{random};

(4) Calculate the total cost S:

$$S = dist(p, o_{random}) - dist(p, o_j) \quad (1)$$

(5) If the total cost $S < 0$, indicates that the non-representative object o_{random} is the optimal solution, the element o_{random} can replace the element o_j to form a new set of k representative objects and continue to return to step (2) to do a new round of object assignment;

(6) If the total cost $S > 0$, indicating that the representative object o_j is better solution, go to step (3), re-select the non-representative object of the total cost of comparison[8], until the total cost S does not change, the total cost of the smallest k clusters, Go to step (7);

(7) Reverse learning the original k clusters obtained in step (6) to obtain k corresponding reverse clusters;

(8) The k clusters obtained by the typical PAM clustering algorithm are arranged and combined to obtain k reverse clusters by reverse learning to obtain $k \times k$ cluster clusters;

(9) Calculate the inter-cluster spacing $a(o)$, inter-cluster spacing $b(o)$ and contour coefficient $s(o)$, compare $s_1, s_2, ..., s_{k \times k}$ and find the cluster combination with the largest contour coefficient.

3. Clustering Algorithm Based on Density Peaks

In 2014, Alex Rodriguez et al. Proposed a new density-based density peak clustering (DPC) algorithm in Science[9]. The main steps of the DPC algorithm are to assume that the centers of the clusters are all densities lower than their centers, and the distances of these points are the lowest compared to other cluster centers. According to the decision-making map, the density peak is selected

as the initial cluster center and the remaining data points are allocated to obtain the clustering result[10].

According to the inspiration of density peak clustering, combining the initial clustering center of density peak with the improved OPAM algorithm can improve the clustering effect and shorten the computing time. Algorithm operation idea: In the data set, the clustering centers are surrounded by the neighbors with lower local density[11], and these low local density points are located at a greater distance from other higher-level local density points.

Let the data points of the sample be i, the local density is ρ_i, and the local density ρ_i of the data points i is calculated as:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{2}$$

Where, $\chi(x) = \begin{cases} 1 & if\ x < 0 \\ 0 & otherwise \end{cases}$, d_c is the truncation distance, for large amounts of data, the local density is essentially the relative density between data points[12], so d_c is robust. The definition δ_i is the minimum distance of a data point i from any point of higher density:

$$\delta_i = \min_{j:\rho_j > \rho_i}(d_{ij}) \tag{3}$$

For the point of maximum local density, the need for special treatment, the general point value is changed:

$$\delta_i = \max_j(d_{ij}) \tag{4}$$

4. OPAM Clustering Algorithm Based on Density Peak Optimization Initial Center

4.1 Algorithm Ideas

In this paper, OPAM clustering algorithm based on density peak optimization initial center, namely DP-OPAM algorithm, is proposed. It is difficult to set the initial center of OPAM algorithm. If the initial value is less deviated from the final center point, the convergence speed is faster[13]. If the deviation of the initial value from the final center point is large, Slow, and easy to fall into the local optimal problems. Through density peak optimization, clusters of data sets are identified, and a reasonable initial cluster center of the data set is selected[14], which effectively reduces the number of clustering iterations, shortens the clustering time and improves the clustering accuracy.

4.2 Algorithm Flow

DP-OPAM algorithm specific steps are as follows:

Step 1 Initialization

(1) Initialize the distance matrix $D=\{d_{ij}\}, i, j = 1,...,n$, for each data point and determine the cutoff distance d_c.

(2) Calculate the local density according to formula (2) and calculate the high density distance δ_i of sample using formula (3).

(3) Construct the decision-making map of horizontal axis ρ and vertical axis δ as the initial cluster center by selecting the data points with high local density ρ and high density δ, and the peak points far from the upper right corner of most samples as the initial clustering center [15].The number of peak points is the number of clusters k.

Step 2 Construct the initial cluster center

Calculate the minimum distance between each point in the dataset and each cluster center, and allocate the remaining sample points to the center of the nearest initial cluster to form the initial segmentation and calculate the sum of squares of clustering errors[16].

Step 3 Reverse Learning and Substitution into PAM Algorithm

(1) Reverse learning the k original clusters obtained above to obtain k corresponding reverse clusters.

(2) The k clusters obtained by the typical PAM clustering algorithm are arranged and combined by reverse learning to obtain k reverse clusters, and $k \times k$ cluster combinations is obtained.

(3) Calculate the intra-cluster spacing $a(o)$, the inter-cluster spacing $b(o)$ and contour coefficient $s(o)$,compare $s_1, s_2, ..., s_{k \times k}$ and find the cluster combination with the largest contour coefficient.

5. Experiment Analysis

In order to verify the DP-OPAM algorithm proposed in this paper, we use the UCI dataset and the collected dataset of a certain area traffic to verify. The experimental environment is Inter®CoreTM i3-2300M CPU @ 2.2GHZ, 4GB, 500GB hard drive, Matlab2013a application software.

Table 1 UCI Data Sets

Name	Numble	Attributes	Cluster
Iris	150	4	3
Wine	178	13	3
Glass	214	9	7

Use the dataset and traffic data in Table 1 to evaluate the experimental results. According to OPAM Clustering Algorithm Based on Density Peak Optimization Initial Center, the algorithm time has been improved, operating efficiency increased. Table 2 reflects the two clustering algorithm running time comparison. Meanwhile, the contour coefficient of DP-OPAM clustering algorithm is reflected in Fig. 1. The cluster center of OPAM clustering algorithm combined with sample collection is similar to the cluster center used in small dataset clustering, and can partially replace the whole. The DP-OPAM algorithm can also improve the OPAM algorithm based on sample collection to find clusters with better clustering. Also shows that it can be applied to large data sets.

Table 2 Comparison of runtime of two algorithms on dataset

Name	OPAM	DP-OPAM
Iris	0.2144	0.2103
Wine	0.2235	0.2211
Glass	0.2201	0.2198

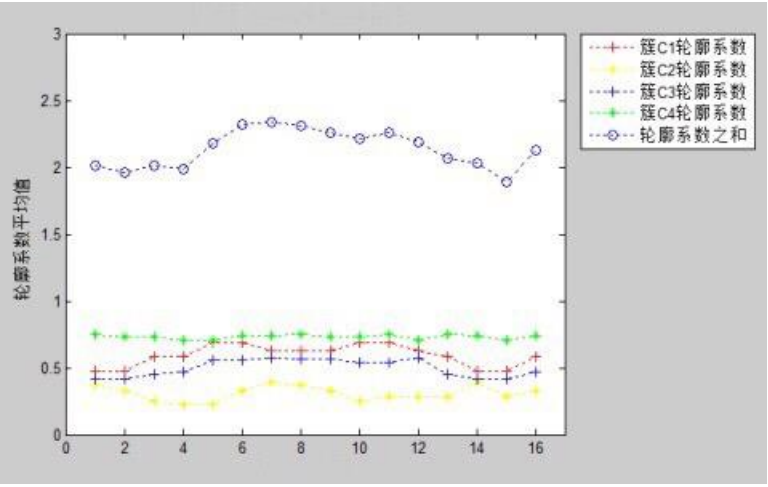

Fig. 1 DP-OPAM Cluster assessment chart

6. Conclusion

In this paper, the OPAM algorithm based on the initial peak of the density peak is used. Based on the original algorithm, the density peak cluster is added to solve the problem that the initial center is difficult to set in the OPAM algorithm and the fluctuation is greatly affected, and the operation efficiency of the algorithm is improved. The next step will be to further optimize the algorithm, improve the accuracy of the algorithm and enhance the robustness.

References

[1] Ester M, Kriegel H P, Sander J, et al.A density-based algorithm for discovering clusters in large spatial database with noise[C]//KDD-96:Proceedings of the 2nd International conference on Knowledge Discovering and Data Mining. Portland: Oregon, 1996: 226-231.
[2] Rodriguez A, Laio A. Machine learning. Clustering by fast search and find of density peaks.[J]. Science, 2014, 344(6191):1492.
[3] Wang S, Wang D, Caoyuan L I, et al. Clustering by Fast Search and Find of Density Peaks with Data Field[J]. Chinese Journal of Electronics, 2016, 25(3):397-402.
[4] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases[C]// ACM SIGMOD International Conference on Management of Data. ACM, 1996: 103-114.
[5] Mart&xed, nezP&xe, Rez &, et al. A density-sensitive hierarchical clustering method [J]. Psychometrika, 2012, 32(3): 241-254.
[6] He Zhenfeng.A restriction-based PAM algorithm [J] .Computer Engineering and Applications, 2006, 42 (6): 190-192.
[7] Tizhoosh H R. Opposition-Based Learning: A New Scheme for Machine Intelligence[C]// Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on. IEEE Xplore, 2005: 695-701.
[8] Park H S,Jun C H.A Simple and Fast Algorithm for K-medoids Clustering[J].Expert Systems with Applications,2009,36(2):3336-3341
[9] Dai Jiao, Zhang Mingxin, Zheng Jinlong, et al.Optimization of fast clustering algorithm based on density peak [J] .Computer Engineering and Design, 2016, 37 (11): 2979-2984.
[10] Liu Cangsheng, Xu Qinglin.A fuzzy C-means clustering algorithm based on density peak optimization, October 12, 2017 [J] .Computer Engineering and Applications, 2017.
[11] Tian Shixiao, Ding Lixin, Zheng Jinqiu.K-means Text Clustering Algorithm Based on Density Peak Optimization [J] .Computer Engineering and Design, 2017,38 (4): 1019-1023.
[12] Bie R, Mehmood R, Ruan S, et al. Adaptive fuzzy clustering by fast search and find of density peaks[J]. Personal & Ubiquitous Computing, 2016, 20(5):785-793.
[13] Shen Y C, Zhang H. Automatically Selecting Cluster Centers in Clustering by Fast Search and Find of Density Peaks with Data Field[C]// International Conference on Information Systems Engineering. IEEE Computer Society, 2017:32-36.
[14] Ruan S, Mehmood R, Daud A, et al. An Adaptive Method for Clustering by Fast Search-and-Find of Density Peaks: Adaptive-DP[C]// WWW. 2017.
[15] Mehmood R, Bie R, Dawood H, et al. Fuzzy Clustering by Fast Search and Find of Density Peaks[C]// International Conference on Identification, Information, and Knowledge in the Internet of Things. IEEE, 2016:785-793.
[16] Ester M, Kriegel H P, Xu X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]// International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996:226-231.