An improved density peaks clustering algorithm based on CURE

Baiyan Chen[a,*], Kai Zhou[b]

School of Computer & Software, Nanjing University of Information Science & Technology, Jiangsu Nanjing, China
[a]1170935526@qq.com, [b]992469196@qq.com
**Corresponding author*

Abstract: *As a new density-based clustering algorithm, clustering by fast search and find of Density Peaks (DP) algorithm regards each density peak as a potential clustering center when dealing with a single cluster with multiple density peaks, therefore it is difficult to determine the correct number of clusters in the data set. To solve this problem, a mixed density peak clustering algorithm namely C-DP was proposed. Firstly, the density peak points were considered as the initial clustering centers and the dataset was divided into sub-clusters. Then, learned from the Clustering Using Representatives algorithm (CURE), the scattered representative points were selected from the sub-clusters, the clusters of the representative point pairs with the smallest distance were merged, and a parameter contraction factor was introduced to control the shape of the clusters. The experimental results show that the C-DP algorithm has better clustering effect than the DP algorithm. The comparison of the F-measure Index shows that the C-DP algorithm improves the accuracy of clustering when datasets contain multiple density peaks in a single cluster.*

Keywords: *Density Peak, Hierarchical Clustering, Cluster Merging, Representative Point, Contraction Factor.*

1. Introduction

Clustering is an unsupervised learning method which organizes data into different clusters to find the inherent hidden patterns of data[1]. It is widely used in many fields, such as image processing [2], pattern recognition [3], bioinformatics [4], microarray analysis [5] and social network [6]. Clustering algorithm can gather more similar data into the same cluster, which is widely used in the field of data mining [7].Now many clustering algorithms based on different features have been proposed [8]. Among them, the density-based clustering algorithm is popular among researchers. Because even in the presence of noise, it can find clusters of arbitrary shapes in large data sets [9].Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [10] is an advanced density based clustering algorithm, which uses the minimum neighborhood of data to find clusters of arbitrary shape. However, the effectiveness of DBSCAN is easily affected by input parameters [11].

A clustering by fast search and find of density peaks (DP) algorithm was proposed in reference [12]. The DP algorithm selects the clustering center according to the peak density, so it has a high operating efficiency and requires less manual intervention. However, when the single data set processed by the DP algorithm contains multiple density peaks, its defects will be exposed. Because the DP algorithm uses different density peaks as potential clustering centers, this causes the original cluster to be divided into multiple clusters, which ultimately makes the clustering effect unsatisfactory.

To solve this problem, this paper proposes a new mixed density peaks clustering (C-DP) algorithm. Firstly, DP algorithm is used to find the initial cluster center. Then, we use the Clustering Using Representatives (CURE) algorithm to select the representative points of the cluster from the initial clustering. In order to improve the quality of clustering, the clusters of representative point pairs with minimum distance are merged. The experimental results show that the C-DP algorithm can recognize clusters of arbitrary shape on UCI data sets.

2. Methodology

2.1. DP algorithm

The DP algorithm assumes that there is a data set S, the local density of the cluster center is higher, and the local density of neighboring points is lower. At the same time, points with low local density are far away from any points with higher local density. At this time, the DP algorithm must calculate two key parameters on each data point χ_i in the data set S. One is the local density ρ_i, the other is the distance δ_i from the high-density point.

Definition 1 In data set S, the local density ρ_i of data point χ_i is expressed as the number of data points whose distance from χ_i is less than d_c (excluding χ_i itself), as follows:

$$\rho_i = \sum_j \chi_i(d_{ij} - d_c) \tag{1}$$

Where:

$$\chi(x) = \begin{cases} 1, & x < 0 \\ 0, & \text{other} \end{cases} \tag{2}$$

$$d_{ij} = dist(\chi_i, \chi_j) \tag{3}$$

Where d_{ij} is the distance between data points χ_i and χ_j, which is calculated based on Euclidean distance.

The parameter $d_c > 0$ is the cutoff distance, and the threshold need set. According to experience, d_c can be selected to make the average number of adjacent points about 1% to 2% of the total points in the dataset.

Definition 2 δ_i is the minimum distance from the data point χ_i to any point with higher density.The calculation method is as follows:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i}(d_{ij}), & i \geq 1 \\ \max\limits_{j}(d_{ij}), & \rho_i \text{ is the highest global density} \end{cases} \tag{4}$$

When the local density of χ_i is the largest, δ_i represents the distance between χ_i and farthest points in the data set. When χ_i does not have the maximum local density, δ_i represents the distance between χ_i and the closest point among all the points whose local density is greater than χ_i.

The specific steps of the DP clustering algorithm are as follows:

Step 1 Initialize the data, calculate d_{ij} by formula (3), arrange in ascending order and find the value from 1% to 2% in ascending order to determine the cutoff distance d_c.

Step 2 Calculate the local density ρ_i and high-density distance δ_i of each data point χ_i by formulas (1) and (2) respectively.

Step 3 Construct a decision graph with ρ as the horizontal axis and δ as the vertical axis. The data points with larger local density ρ and high-density distance δ are selected as the cluster center according to the decision graph.

Step 4 Calculate the minimum distance between each point and cluster center, then assign each data point to the closest cluster center.

Step 5 Filter the noise outlier data so that the density does not exceed the boundary.

Step 6 Complete clustering and get cluster labels.

2.2. CURE algorithm

The CURE algorithm is a novel hierarchical clustering algorithm. Instead of using a single centroid or object to represent a cluster, a fixed number of representative points in the data space are selected.The method of generating representative points of a cluster is as follows: Firstly, select the scattered objects in the cluster, then move them according to the shrinkage factor [13].

Definition 3 For a data set S, C_i represents a cluster in the data set S, m_i represents the center point of C_i, and p_i represents a set of representative points of the cluster C_i.

$$dist(C_i, C_j) = \min_{\chi_i \in \rho_i, \chi_j \in \rho_j} dist(\chi_i, \chi_j) \tag{5}$$

The $dist(C_i, C_j)$ is calculated based on Euclidean distance. At each step of the algorithm, two clusters of representative point pairs with the closest distance are merged, where each point comes from a different cluster.The number of data items of C_i can be represented by $|C_i|$. After the two clusters of C_i and C_j are merged, the new center point $m.mean$ is calculated as follows:

$$m.mean = (|C_i| m_i + |C_j| m_j) / (|C_i| + |C_j|) \tag{6}$$

The number of new merged data items is denoted by p, and the new merged representative point $m.rep$ is calculated as follows:

$$m.rep = p + \alpha^*(m.mean - p) \tag{7}$$

Among them, α is the shrinkage factor. The CURE algorithm is sensitive to the shrinkage factor α, which is mainly used to adjust the shape of the cluster. When $\alpha = 1$, the CURE algorithm is equivalent to a hierarchical clustering algorithm based on the centroid representative cluster; when $\alpha = 0$, the CURE algorithm is equivalent to use the minimum spanning tree clustering algorithm.

The CURE algorithm uses a set of points to represent a cluster, and controls the shape of the cluster through a shrinkage factor, which helps to control the influence of isolated points. Therefore, the CURE algorithm is more robust to the processing of isolated points, and can effectively identify clusters with relatively large shape changes.

2.3. C-DP algorithm

Although, the DP algorithm performs well in determining the density and number of cluster centers. However, when the DP algorithm processes a single data set containing multiple density peaks, the peaks of different densities are regarded as potential cluster centers. At this time, the original cluster will be divided into multiple clusters, which will eventually lead to poor clustering results. Compared with the DP algorithm, the C-DP algorithm uses the DP algorithm to find the initial cluster, introduces a shrinkage factor α and selects the representative points in the initial cluster. Then the algorithm uses the shrinkage factor to move the representative point to achieve the purpose of controlling the shape of the cluster. Here, the C-DP algorithm uses the minimum distance measurement method as a measure of similarity between clusters, and two clusters with the closest distance between different clusters representing a point pair will be merged.

In the data set, a single cluster may contain multiple density peaks. The C-DP algorithm takes into account this situation and draws on the merging strategy of hierarchical clustering based on the initial clustering of the DP algorithm. The C-DP algorithm uses several scattered data points as representative points, and introduces a parameter shrinkage factor α to adjust the shape of the cluster and improve the efficiency of clustering. The algorithm flow chart is shown in Figure 1.

The algorithm steps are as follows:

Step 1 Initialize the data and determining the cutoff distance d_c.

Step 2 Calculate the local density ρ_i of each data point χ_i and the distance δ_i from the high-density point by formulas (1) and (2), and then determine the initial cluster center and number through the decision diagram.

Step 3 Take the initial cluster center as the center point and substitute it into formula (6), where the number of merged data items p is zero. Finally, a set of points is obtained as the representative points of the cluster and a representative point set p_i.

Step 4 Specifying the shrinkage factor α value.

Step 5 The distance between the clusters is calculated by formula (5), and the two clusters with the smallest distance are merged to form a new cluster.

Step 6 Calculate the center point and representative point of the new cluster by formulas (6) and (7), and then go to step 5.

Step 7 Calculate the minimum distance between each point and cluster center, then assign each data point to the closest cluster center.

Step 8 Filter the noise outlier data so that the density does not exceed the boundary.

Step 9 Complete clustering and get cluster labels.

It is worth noting that the shrinkage factor α value needs to be set according to the specific data set, and its value directly affects the effect of clustering.

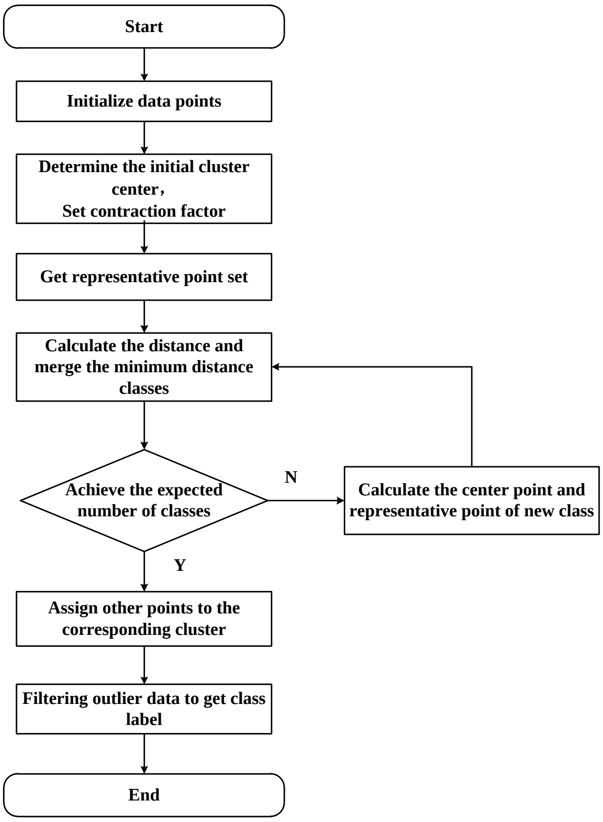

Start
Initialize data points
Determine the initial cluster center，
Set contraction factor
Get representative point set
Calculate the distance and merge the minimum distance classes
Achieve the expected number of classes
N
Calculate the center point and representative point of new class
Y
Assign other points to the corresponding cluster
Filtering outlier data to get class label
End

Figure 1: C-DP algorithm flow chart.

3. Results and discussion

3.1. Experiment description

The effect of clustering is compared through the visualization graph on the UCI data set; the accuracy of clustering results is measured by F-measure Index.

In the experiment, the parameter d_c is between 1% and 2%, and the contraction factor α is between 0.2 and 0.7. According to the results of several experiments, the threshold is adjusted. The contraction factor is a heuristic parameter, which is defined by the user. The optimal experimental effect is obtained by adjusting the parameter values through experiments. The experimental results are classified by the color area of the points. The data set used in the experiment is shown in Table 1.

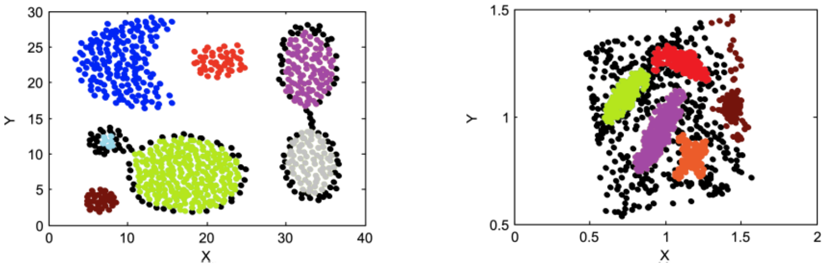
Table 1: Experimental data.

Data sets	Number of clusters	Number of instances	dimension
Iris	3	150	4
Seeds	3	210	7
Wine	3	178	13
Glass	6	214	9
S1	15	5000	2
Spiral	3	312	2
Aggregation	7	788	2
4k2_far	4	400	2

The experimental environment is windows 10 64 bit operating system, Intel Core i7 @ 2.7 GHz CPU, 4 GB memory, and the software is matlab 2014a.

3.2. *Experimental comparison*

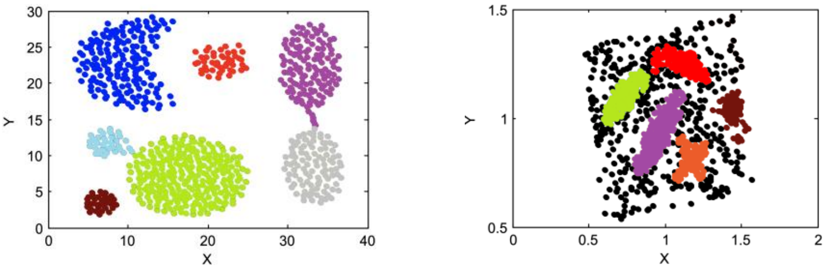
Figure 2 shows the clustering results of DP algorithm on UCI datasets. Obviously, multiple density peaks are found in data set S1 and Aggregation, and DP algorithm will divide the cluster into several ones, which leads to the poor clustering accuracy of DP algorithm.



(a) Aggregation (b) S1

Figure 2: Clustering results of DP algorithm.

Although the initial clustering center strategy selected by C-DP and DP algorithm is the same, after finding the initial clustering center, C-DP algorithm can better complete the clustering task through the appropriate cluster merging strategy.As shown in Figure 3, the clustering result of C-DP algorithm is more accurate and effective through the appropriate cluster merging strategy.



(a) Aggregation (b) S1

Figure 3: Clustering results of C-DP algorithm.

3.3. Algorithm performance comparison

In order to verify the performance of C-DP algorithm, we compare it with other mainstream algorithms, and use F-measure Index to measure the clustering accuracy.

Table 2: The performance comparison.

Data sets	DBSCAN	AP	DP	C-DP
Iris	0.7409	0.4849	0.6704	0.7710
Seeds	0.4001	0.3881	0.8019	0.8171
Wine	0.5829	0.3147	0.5894	0.6501
Glass	0.5049	0.2874	0.5511	0.5527
S1	0.7359	0.2881	0.7624	0.7831
Aggregation	0.7725	0.3386	1	1
Spiral	0.7801	0.2696	1	1
4k2_far	0.7974	0.3271	0.9327	0.9405

Table 2 shows the comparison results of C-DP algorithm, DP algorithm, DBSCAN algorithm and AP algorithm on the UCI data set. Obviously, the C-DP algorithm is better than other algorithms. Moreover, compared with the DP algorithm, the performance of C-DP algorithm is improved, while the clustering effect of AP algorithm is poor. This shows that the C-DP algorithm can effectively deal with multi-density peak clusters, and has the ability to deal with irregular shape data sets.

4. Conclusion

Aiming at the defect that there are multiple density peaks in a single cluster in the data set, the DP algorithm cannot obtain accurate clustering results. We proposes a clustering algorithm of mixed density peaks (C-DP). Inspired by the CURE algorithm, we introduced the shrinkage factor α during the initial clustering process of the DP algorithm. Then we select the representative points of the cluster from the initial cluster and move them according to the shrinkage factor to control the shape of the cluster. In the algorithm, the minimum distance measurement method is selected as the measure of similarity between clusters, and the representing point pairs of two clusters with the closest distance between different clusters will be merged. The experimental results show that the C-DP algorithm improves the accuracy of clustering when datasets contain multiple density peaks in a single cluster.

However, the clustering center of each data set in C-DP algorithm is selected by the user, which is difficult to determine accurately. In the future, it is necessary to extend C-DP algorithm to fully adaptive method.

References

[1] Zhen, C., Jiang, C. (2019) Overview of Data Mining in the Era of Big Data. International Core Journal of Engineering, 5, 136-139.

[2] Yan, M., Chen, L., Peng, L. (2016) Parallel programing templates for remote sensing image processing on GPU architectures: design and implementation. Computing, 98, 7-33.

[3] Liu, S., Zou, Y. (2020) An Improved Hybrid Clustering Algorithm Based on Particle Swarm Optimization and K-means. IOP Conference Series: Materials Science and Engineering, 750, 152-158.

[4] Zhao, L., Liu, Z., Levy, S.F. (2018) Bartender: a fast and accurate clustering algorithm to count barcode reads. Bioinformatics, 34, 739-747.

[5] Jothi, R., Mohanty, S.K., Ojha, A. (2019) DK-means: a deterministic K-means clustering algorithm for gene expression analysis. Pattern Analysis and Applications, 22, 649-667.

[6] Zhang, P., Shen, Q. (2018) Fuzzy c-means based coincidental link filtering in support of inferring social networks from spatiotemporal data streams. Soft Computing, 22, 1-11.

[7] Zou H. (2020) Clustering Algorithm and Its Application in Data Mining. Wireless Personal Communications, 110, 21-30.

[8] Gob, N., Rathinavelu A. (2018) Analyzing cloud based reviews for product ranking using feature based clustering algorithm. Cluster Computing, 22, 6977-6984.

[9] Chen, J., Chen, J., Yang D. (2018) A k-Deviation Density Based Clustering Algorithm. Mathematical Problems in Engineering, 2, 1-16.

[10] Liu, S.F., Meng, D.X., Wang X.Y. (2014) DBSCAN algorithm based on grid cell. Journal of Jilin University, 44, 1135-1139.

[11] Karami, A., Johansson, R. (2014) Choosing DBSCAN Parameters Automatically using Differential Evolution. International Journal of Computer Applications, 91, 1-11.

[12] Rodriguez, A., Laio, A. (2014) Clustering by fast search and find of density peaks. Science, 344, 1492-1496.

[13] Kirtee. Panwar. Alka. (2016) Modified CURE algorithm with enhancement to identify number of clusters. International journal of artificial intelligence and soft computing: IJAISC, 5, 226-240.