Mass-Based Density Peaks Clustering Algorithm

Ding Ling, Xu Xiao

Mass-based Density Peaks Clustering Algorithm

Ding Ling[1], Xu Xiao[2,*]

[1] School of computing technology and gaming development, Asia Pacific University of Technology and Innovation, Petaling KL, 57000, Malaysia
tp033295@mail.apu.edu.my
[2] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China
xu_xiao@cumt.edu.cn

ABSTRACT. Density peaks clustering algorithm (DPC) relies on local-density and relative-distance of dataset to find cluster centers. However, the calculation of these attributes is based on Euclidean distance simply, and DPC is not satisfactory when dataset's density is uneven or dimension is higher. In addition, parameter d_c only considers the global distribution of the dataset, a little change of d_c has a great influence on small-scale dataset clustering. Aiming at these drawbacks, this paper proposes a mass-based density peaks clustering algorithm (MDPC). MDPC introduces a mass-based similarity measure method to calculate the new similarity matrix. After that, K-nearest neighbour information of the data is obtained according to the new similarity matrix, and then MDPC redefines the local density based on the K-nearest neighbour information. Experimental results show that MDPC is superior to DPC, and satisfied on datasets with uneven density and higher dimensions, which also avoids the influence of d_c on the small-scale datasets.

KEYWORDS: DPC algorithm, mass-based similarity measure, decision graph, uneven density, higher dimensions

adfa, p. 1, 2011.

1 Introduction

Clustering is named unsupervised learning as it does not depend on the predefinition of classes and the labelling of data samples, and it is an effective technique for data mining [1]. As so far, clustering is applied to the pattern recognition, image processing, genetic research, and many other fields [2].

The main idea of clustering is to classify data objects into multiple clusters according to a measure of similarity. As far as possible, the similarity of the data objects in the same cluster is greater, and the similarity of data objects between different clusters is smaller [3]. Moreover, different clustering targets correspond to different clustering algorithms and the current clustering algorithms are mainly divided into: partition-based clustering, density-based clustering, grid-based clustering, hierarchical clustering and model-based clustering [4]. These different algorithms are suitable for different types of datasets with different advantages and disadvantages.

In 2014, Rodriguez et al. proposed a clustering by fast search and find of density peaks algorithm (DPC) [5]. DPC algorithm uses the local density and relative distance properties of the data to determine the cluster centers quickly and can be used for arbitrary shape datasets and perform sample points allocation effectively [6]. However, it has the following limitations: (1) the calculation of similarity between data samples relies on Euclidean distance simply, which makes DPC cannot get satisfactory clustering results when the data distribution is uneven or the data dimension is higher [7]. (2) A little change of parameter d_c in small-scale datasets will affect the clustering results obviously [8]. At present, many scholars have optimized DPC. Du et al. [9] and Xie et al. [10] both introduced K-nearest neighbours algorithm and considered the local distribution of datasets to redefine the local density, thereby unifying local metrics to reduce the impact of d_c on clustering results. But k nearest neighbours are found still based on the Euclidean distance, which is also unsatisfactory.

To alleviate the adverse influence of the limitation of DPC, this paper proposes a mass-based density peaks clustering algorithm (MDPC). The main innovations of MDPC algorithm include: (1) Consider the environment around the datasets and using mass-based measure to replace the Euclidean distance for measuring the similarity between datasets to improve the clustering accuracy of data with higher dimensions or uneven distribution; (2) Redefine the local density by the improved K-nearest neighbour information of samples and make MDPC independent of d_c.

The remaining parts of this paper are as follows: Section 2, the basic principle of density peaks clustering algorithm and mass-based similarity measure method. In section 3, this paper proposes a mass-based density peaks clustering algorithm and analyses its performance from the theoretical level. Section 4 designs experiments to test this algorithm and other clustering algorithms for comparison on different datasets. Finally, the work done in this paper is summarized and the direction of the next research is given.

2 Related works

2.1 Density peaks clustering algorithm

A density peaks clustering algorithm (DPC) was proposed to find cluster centers fast by Rodrigue and Laio. DPC algorithm is based on an important assumption that the local density of the cluster centers is greater than the local density of the surrounding neighbours and the distance between cluster centers and the points with higher local density is relatively far [11].

DPC algorithm first calculates the local density and relative distance attributes of each data point. The local density is defined as:

$$\begin{aligned} \rho_i &= \sum_j \chi\left(d_{ij} - d_c\right), \\ \chi(x) &= \begin{cases} 1, x < 0, \\ 0, x \geq 0, \end{cases} \end{aligned} \tag{1}$$

Where d_{ij} is the distance between the data points x_i and x_j. d_c is the only input parameter that represents the cut-off distance, and is defined as the average number of neighbours which is around 1% to 2% of the total number of points in the dataset. The relative distance δ_i of the data point x_i is the minimum value of the distance from the point to all points whose local density is larger, and its formula is:

$$\delta_i = \min_{j:\rho_j > \rho_i}\left(d_{ij}\right), \tag{2}$$

For the densest point, we can get:

$$\delta_i = \max_j\left(d_{ij}\right), \tag{3}$$

DPC algorithm selects data points with large ρ_i and δ_i as cluster centers. After DPC determines the cluster centers, it needs to allocate the remaining points to the corresponding clusters. DPC algorithm first assigns all remaining points to its nearest point's cluster whose local density equal to or higher than the current point. Then, a boundary threshold is defined for each cluster to remove noise points.

DPC algorithm is simple and effective, can deal with noise outliers as well as get clusters of arbitrary shape clustering [12]. But, the disadvantages of DPC are obviously: First, the calculation of local density and relative distance is based on the similarity between data nodes, and the measure of similarity simply depends on the Euclidean distance, which causes DPC cannot obtain satisfactory clustering results on complex data, especially when the data distribution is uneven and the data dimension is higher [13]; Second, the calculation of local density depends on the choice of the cut-off distance d_c, but it only considers the global distribution of the data and ignores the local information, which will lead to a big influence of d_c's change on the clustering results, especially on small-scale datasets [14].

2.2 Mass-based similarity measure

Since 1970s, psychologists have stated that the similarities between two instances cannot be simply characterized by geometric models, and the measure of similarity is influenced by the background and the neighbours [15]. Based on this fact, it can define a more appropriate measure of similarity, here called mass-based similarity measure [16].

The basic idea of the mass-based similarity measure is that the two instances of the dense region have similarities less than two instances of the same interval but in the low-density region [17]. The geometric model-based similarity calculation only depends on the geometric position derivation. On the contrary, mass-based measure of similarity mainly depends on the data distribution, that is, the probability mass covering the smallest region of two instances [18].

Assume that D represents a data sample in the probability density function F, and $H \in \mathcal{H}(D)$ represents a hierarchical division that divides the space into non-overlapping and non-empty domains. $R(x, y \mid H; D)$ denotes the minimum domain for H and D over x and y. Notice that $R(x, y \mid H; D)$ is the smallest area covering x and y, similar to the shortest distance in the x and y geometric models.

Mass-based similarity measure defines two parameters t and ψ to represent the number of "iTrees" and the size of each "iTree", and the height of each "iTree" is up to $h = \lceil \log_2 \psi \rceil$. First, build an "iForest" consisting of t "iTree" as the partition structure R. Each iTree is built separately using subset $\mathcal{D} \subset D$, where $| \mathcal{D} | = \psi$. Axis-parallel segmentation algorithm is used at each internal node of the "iTree" to divide the samples at the node into two non-empty subsets until each points are quarantined or reach the maximum height h. After "iForest" is established, all instances in D are traversed to record the mass for each node. The second step is to value the mass. The evaluation through each "iTree" analytical test points x and y, calculate the sum of the mass containing the lowest node of both x and y, that is $\sum_i | R(x, y \mid H_i) |$. Finally, $m_e(x, y)$ is the mean of these mass:

$$m_e(x, y) = \frac{1}{t} \sum_{i=1}^{t} \frac{| R(x, y \mid H_i) |}{| D |} . \tag{4}$$

3 Mass-based Density Peaks Clustering Algorithm

This paper proposes a mass-based density peaks clustering algorithm (MDPC). MDPC algorithm will maintain the central idea of DPC, and quickly find the cluster centers whose local density and relative distance properties are larger, but similarity calculations and local density measurement will be improved.

First, the similarity measure between samples. A mass-based similarity measure will be introduced to replace the Euclidean distance. A new similarity matrix will be derived from Eq.(4).

Then based on the new similarity matrix, the K-nearest neighbours of the sample are found and defined. New local density is defined as:

$$\rho_i = \sum_{j \in KNN(i)} \exp(-m_e(x_i, x_j)) , \tag{5}$$

Where $KNN(i)$ is the k nearest neighbours of point x_i . At the same time, the relative distance attribute of the data sample no longer depends on the similarity of the geometric distance metric, but uses the similarity calculated by equation (4):

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} \left(m_e(x_i, x_j) \right), \ if \ \exists j \ s.t. \rho_i > \rho_j \\ \max\limits_{j} (m_e(x_i, x_j)), \ otherwise \end{cases} . \tag{6}$$

Specific steps of MDPC algorithm are described as algorithm 1.

Algorithm1. MDPC algorithm.

Input: datasets $\boldsymbol{X}$; number of iTree t ; size of each iTree ψ ; number of nearest neighbor k ;

Output: clustering result $\boldsymbol{Y}$.

Step1: Divide the dataset $\boldsymbol{X}$ into t sets of size ψ by random sampling;

Step2: Axis-parallel segmentation is performed for each set to form one “iTree”, and t “iTree” constitute one “iForest”;

Step3: Go through “iForest” and calculate the sample similarity matrix based on the mass-based similarity measure (4).

Step4: Calculate the ρ_i and δ_i of each sample according to the formula (5) and (6) ;

Step5: Select the cluster centers automatically based on the decision graph;

Step6: Assign the remaining data points in the dataset to the nearest point where the density is equal to or higher than the "current point";

Step7: Return the result matrix $\boldsymbol{Y}$.

MDPC algorithm retains the main ideal of DPC algorithm and finds density peaks as cluster centers quickly. However, MDPC algorithm's local density and relative distance properties are measured by the mass-based similarity measure method instead of the simple Euclidean distance, which makes MDPC more efficient in high dimensional datasets and uneven density datasets. In addition, the mass-based similarity between data samples is used to define a new local density based on the improved K-nearest neighbour information. Compared with DPC algorithm, MDPC algorithm

avoids excessive dependence on the d_c, and the local density metric is suitable for any size dataset.

4 Experiments

4.1 Experimental preparation

In order to prove the performance of MDPC algorithm, the experiments were tested on synthetic datasets and real-word datasets. The clustering accuracy *Acc* was used to measure the clustering results. The higher the value of *Acc*, the better the clustering performance of MDPC. If y_i and z_i are the intrinsic class labels and clustering result labels, respectively. $map(\cdot)$ maps each label to a class label by the Hungarian, and the map is optimal. *Acc* is calculated as follows:

$$Acc = \sum_{i=1}^{N} \delta\left(y_i, map\left(z_i\right)\right) \Big/ n \ . \tag{7}$$

In addition to DPC algorithm, we compared the MDPC algorithm with the optimization algorithm DPC-KNN. The parameter d_c in DPC algorithm is in the interval [0.2%-6%], and k in DPC-KNN algorithm is taken from 5 to 7. In MDPC, both t and ψ take the default values 100 and 256. The value of k is also taken from 5 to 7.

4.2 Results and evaluation

Synthetic datasets.

This section conducts MDPC testing on the synthetic dataset D, which is a typical dataset containing three clusters with uneven density. Along with 97 samples, D has two attributes.

The experiment shows the result of the two-dimensional dataset visually. One colour represents one cluster. MDPC algorithm and DPC algorithm are clustered on the above D datasets respectively. The results are shown in Figure 1.

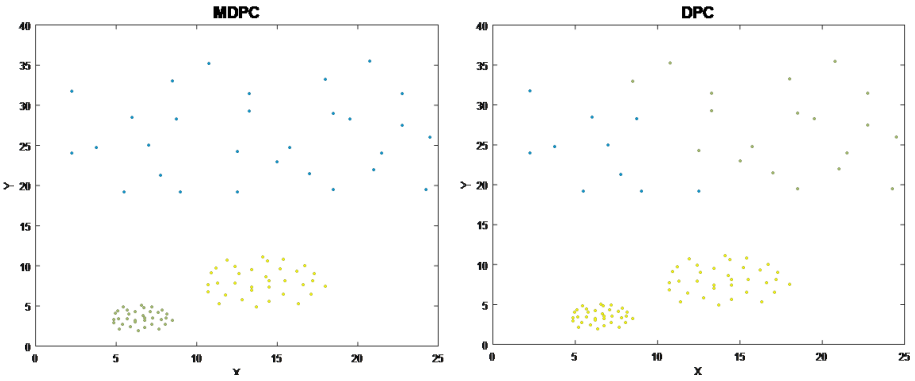

Fig. 1. Clustering results of different algorithms on D dataset

From Figure 1, it can be seen that MDPC can handle datasets with uneven density very well. On dataset D, MDPC algorithm can be well clustered into 3 categories, but DPC does not divide the dataset into 3 classes well, because DPC simply uses the geometric distance between data to measure similarity and calculate the local density and relative distance properties. Therefore, for datasets with uneven dataset distribution, DPC does not recognize all clusters well.

Although MDPC algorithm and DPC algorithm can obtain satisfactory clustering results by selecting suitable parameters on the dataset with relatively uniform distribution. But DPC algorithm needs to select the appropriate d_c. The changes of d_c have great influence on the clustering results. MDPC algorithm no longer need to select d_c. Although there still has one parameter, but since MDPC still selects the cluster centers according to the characteristics of DPC, the local density of the cluster center must be higher. Small changes in k have little effect on the clustering results.

MDPC algorithm introduces mass-based similarity measure method and considers the data distribution environmental, thus MDPC algorithm is more effective than DPC in dealing with uneven density. In addition, MDPC overcomes the influence of d_c on the clustering results based on the K-nearest neighbours algorithm while adding an optional parameter k, but the change of k has a little effect on the clustering results.

Real-word datasets.

This section conducts MDPC on 4 real-word datasets. The characteristics of the experimental data are shown in Table 1. As the changes of d_c in DPC algorithm have a greater impact on the clustering results on small-scale datasets, and the clustering results of DPC on the datasets with higher dimensions is not satisfactory. Thus, this experiment selected the classic small-scale datasets and contains higher dimensions.

Table 1. UCI Datasets

Datasets	Samples	Attributes	Categories
Seeds	210	7	3
Wine	178	13	3
WDBC	569	30	2
Soybean	47	35	4

In this experiment, MDPC algorithm, DPC algorithm and DPC-KNN algorithm were clustered in the above 4 datasets respectively. The clustering results are shown in Table 2, and the corresponding optimal parameters are given. Bold is the best result in the algorithms, while MDPC gives 20 test averages.

Table 2. Accuracy of Different Algorithms on Different Datasets

Datasets	MDPC	DPC	DPC-KNN
Seeds	**89.85** ($k=7$)	89.524 ($d_c=1\%$)	89.524 ($k=7$)
Wine	**94.086** ($k=7$)	69.101 ($d_c=0.2\%$)	53.933 ($k=7$)
WDBC	**92.249** ($k=6$)	62.917 ($d_c=2\%$)	79.438 ($k=7$)
Soybean	**100** ($k=6$)	89.362 ($d_c=2\%$)	91.49 ($k=7$)

It can be seen from Table 2 that the overall clustering performance of MDPC algorithm is better than DPC and DPC-KNN. DPC algorithm on the higher-dimensional dataset is not satisfactory, and d_c needs an appropriate choice. Although DPC-KNN algorithm avoids the choice of d_c, the clustering result is not ideal compared with MDPC algorithm. MDPC algorithm uses the mass-based similarity replaces the geometric distance and considers the environment of data distribution to work well for datasets with higher data dimension. In addition, MDPC also chooses k nearest neighbours to measure local density which avoids the selection of d_c. The increased parameter k in MDPC has little effect on the clustering results as the cluster centers in densely dense areas. Thus, MDPC is superior to DPC and DPC-KNN.

5 Conclusions

This paper proposes an optimized density peaks clustering algorithm based on a novel mass-based similarity measure. The mass-based measure is used to calculate the similarity between data samples first, and the obtained similarity is introduced into the K-nearest neighbour information of the samples. A new local density is redefined by the K-nearest neighbour information to avoid the influence of parameter selection, and improves DPC algorithm on the higher-dimensional and uneven-density datasets. In addition, MDPC algorithm matins the main steps of DPC algorithm to select the cluster centers, thus the choice of increased parameter is robust. MDPC algorithm is superior to the traditional DPC algorithm and the optimized DPC-KNN algorithm.

In this paper, how to allocate the no-center points of MDPC algorithm instead of adopting a one-step allocation strategy, and the effective treatment of noise points requires further exploration.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant no.61672522 and no.61379101.

References

1. Morris, K., Mcnicholas, P.: Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures. Computational Statistics & Data Analysis. 97, 133-150 (2016)

2. Ivannikova, E., Park, H., Hämäläinen, T., et al.: Revealing community structures by ensemble clustering using group diffusion. Information Fusion. 42, 24-36 (2018)
3. Slimen, Y., Allio, S., Jacques, J.: Model-based co-clustering for functional data. Neurocomputing. 291, 97-108 (2018)
4. Fraley, C., Raftery, A.: Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association. 97, 611-631 (2011)
5. Rodríguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science. 344, 1492-1496 (2014)
6. Xu, X., Ding, S., Du, M., et al.: DPCG: an efficient density peaks clustering algorithm based on grid. International Journal of Machine Learning & Cybernetics. 9, 743-754 (2018)
7. Ding, S., Du, M., Sun, T., et al.: An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood. Knowledge-Based Systems. 133, 294-313 (2017)
8. Liu, R., Wang, H., Yu, X.: Shared-nearest-neighbor-based clustering by fast search and find of density peaks. Information Sciences. 450, 200-226 (2018)
9. Du, M., Ding, S., Jia, H.: Study on density peaks clustering based on k-nearest neighbors and principal component analysis. Knowledge-Based Systems. 99, 135-145 (2016)
10. Xie, J., Gao, H., Xie, W., et al.: Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K -nearest neighbors. Information Sciences. 354, 19-40 (2016)
11. Shi, Y., Chen, Z., Qi, Z., et al.: A novel clustering-based image segmentation via density peaks algorithm with mid-level feature. Neural Computing and Applications. 28, 29-39 (2017)
12. Bai, L., Cheng, X., Liang, J., et al.: Fast density clustering strategies based on the k-means algorithm. Pattern Recognition. 71, 375-386 (2017)
13. Wang, M., Min, F., Zhang, Z., et al.: Active learning through density clustering. Expert Systems with Applications. 85, 305-317 (2017)
14. Zhou, L., Pei, C.: Delta-distance based clustering with a divide-and-conquer strategy: 3DC clustering. Pattern Recognition Letters. 73, 52-59 (2016)
15. Krumhansl, C.: Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological Review. 85, 445-463.(1987)
16. Kai, M., Zhu, Y., Carman, M., et al.: Overcoming key weaknesses of distance-based neighbourhood methods using a data dependent dissimilarity measure. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'16, San Francisco, 13 - 17, August 2016, California, USA, p. 1205-1214 (2016)
17. Aryal, S., Kai, M., Haffari, G., et al.: Mp-dissimilarity: a data dependent dissimilarity measure. 2014 IEEE International Conference on Data Mining,14-17, December 2014, Shenzhen, China, p. 707-712 (2014)
18. Chen, B., Ting, K., Washio, T., et al.: Half-space mass: a maximally robust and efficient data depth method. Machine Learning. 100, 677-699.(2015)