ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp.81-85
doi: 10.18178/wcse.2017.06.013

A Clustering Algorithm Based on Find Density Peaks

Wang Peng[1+] and Wang Junyi [1]
[1]School of Software, Inner Mongolia University, Inner Mongolia 010021, China

Abstract. A clustering algorithm named "Clustering by fast search and find of density peaks" is for finding the centers of clusters quickly. The algorithm has the advantages of fast clustering speed and simple realization. But after several experiments on the algorithm,we found that the algorithm results excessively depends on the selected cluster center, if a cluster has multiple density peaks or multiple clusters share the same density peak leads to incorrect clustering, and the algorithm will be determined too many points as noise. In view of the disadvantages, a new way proposed to optimization of CFSFDP by using clusters local density distribution graph and hierarchical clustering algorithm,and identify noise according the outlier degree of the point (DMCFSFDP). Firstly, the new algorithm used CFSFDP algorithm to determined cluster centers and clustering data set. Secondly, DMCFSFDP division clusters based on local density distribution graph and merged the clusters that could be merged by using improved hierarchical clustering algorithm. Finally, the algorithm uses the outlier degree to identify the noise. The results of experiments have shown the DMCFSFDP algorithm is more effective than CFSFDP algorithm in clustering.

Keywords: clustering, density peaks, local density distribution, merge clusters, otulier degree

1. Introduction

Clustering is one of most fundamental and important topic in machine learning. For various applications , there are many clustering algorithm and classified into several categories mainly : partition-based algorithm, like K-means[1] and K-medoids[2], Chamelen[3] and CURE[4] are based on hierarchies; density-based algorithm like DBSCAN [5] and OPTICS[6] ; grid-based algorithm like "Wave Cluster"[7]. These algorithms have some problems: K-means and K-medoids are not able to detect nonspherical clusters; Chamelen and Wave cluster are sensitive to the noise; and DBSCAN and OPTICS are not suitable for high dimensional data.

Alex and Alessandro [8] proposed an algorithm that cluster centers were have higher local density and the distance between higher local density points is relatively large. In the process of clustering according to this algorithm, the cluster centers appear intuitively, and clusters are recognized regardless of their shape and dimension. The author proves that the algorithm is effective by experiments, and through the image clustering to prove that the algorithm can handle high-dimensional data. In addition, some scholars made a thorough study on the algorithm; they make the algorithm more accurate by selecting the optimal threshold, and by combining other algorithms to make the algorithm more robust. However, the algorithm still has some problems in some cases. Through some experiment in the context of this algorithm (These experiments will be given in following section 2), we found that the cluster centers has a great impact on the clustering results. For some data sets, if a cluster has multiple densities peaks, the algorithm will excessive division the cluster, and if multiple clusters share the same density peak, the algorithm will merge clusters with large differences in density, and the algorithm often determines too many data points as noise. Thus, it becomes a bottleneck for this algorithm to recognize any shape and density clusters.

To break through the bottleneck, in this paper, we propose a method to solve the above problems. We optimize the clustering results by describing the local density distribution of the data set, and calculated outlier degree of the data points to identify the noise.

2. Algorithm principle

[+] Corresponding author. Tel.: +18665748660.
E-mail address: 58040050@163.com.

2.1. CFSFDP(Clustering by fast search and find of density peaks)

For a data set D, the algorithm computes two characteristics for each point in this data set, local density ρ_i and the distance δi from the higher local density points. The local density is calculated by Gaussian kernel or cut kernel, the local density is defined as Eq. (2-1).

$$\rho_i = \sum_{j \neq i} \exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \tag{2-1}$$

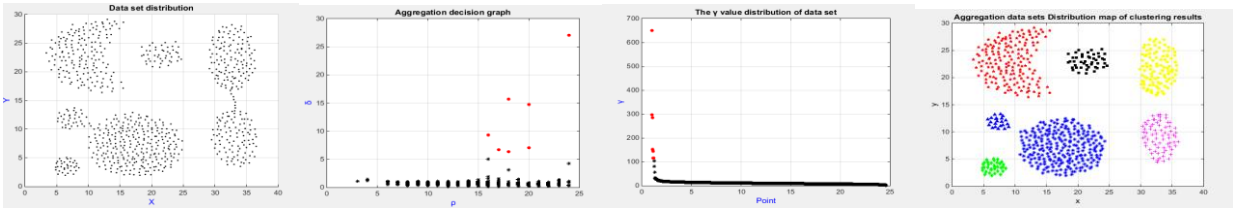
Where the d_{ij} is the distance between two points, and the d_c is threshold, it is determined by empirical experience. As the author suggests, we can choose d_c so that the average number of neighbors is around 1 % to 2% of the total number of points in the data set. δ_i is defined by finding the distance from the nearest point with a higher local density. For the point has the highest local density, δ_i is the maximum distance to other points. δ_i is defined as Eq.(2-2).

$$\delta_i = \begin{cases} \min\limits_{\rho_j > \rho_i}(d_{ij}) & \exists \rho_i < \rho_j \\ \max\limits_{j=1,2,3\ldots n}(d_{ij}) & \text{otherwise} \end{cases} \tag{2-2}$$

The algorithm got the cluster centers by drawing the decision graph, and the cluster centers are defined as that has high ρ_i and δ_i simultaneously, For example, Fig.1 (b) shows the decision graph for the Aggregation [9] data set, the abscissa is ρ_i and the ordinate is δ_i, we marked the cluster center by red color in Fig.1 (b). The author mentions a heuristic approach in the article to find cluster center, introduced a parameter γ which is also considered ρ_i and δ_i of the data point, γ is defined as Eq. (2-3).

$$\gamma_i = \rho_i \times \delta_i \tag{2-3}$$

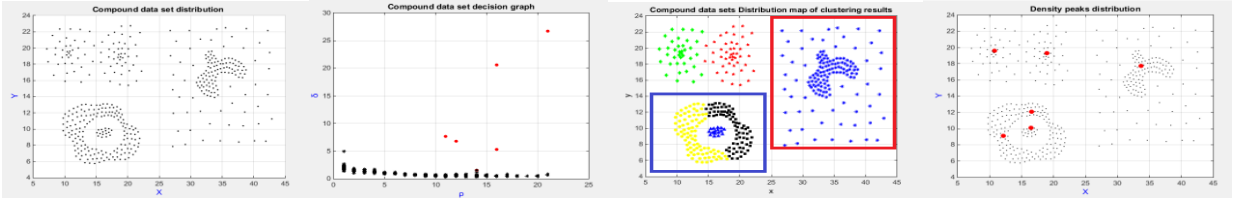
By sorting the γ of all data points in descending order, Fig.1(c) shows the γ distribution of the Aggregation data set. We can clearly find that the cluster centers has a higher γ (marked by red color) , after the cluster centers are determined, the rest points is assigned to the same cluster as its nearest neighbor of higher density. For the noises, we first find the maximum local density of the intersecting regions of each clusters, and the point smaller than the density is determined as noise. Fig.1 (d) shows the clustering result of Aggregation data set by the algorithm, the data set has seven clusters with different shapes. Points with the same color and shape belong to same cluster.



(a) Data set distribution (b) Decision graph (c) γ distribution (d) Clustering result

Fig.1: The CFSFDP algorithm is examples on the Aggregation data set

However, the clustering result is affected by the cluster centers very much. As the clustering result shows that each cluster has only one density peak, and if a cluster has multiple density peaks or multiple clusters share the same density peak will make the clustering result incorrect. Fig.2 shows the clustering process of Compound [10] data set by CFSFDP algorithm.



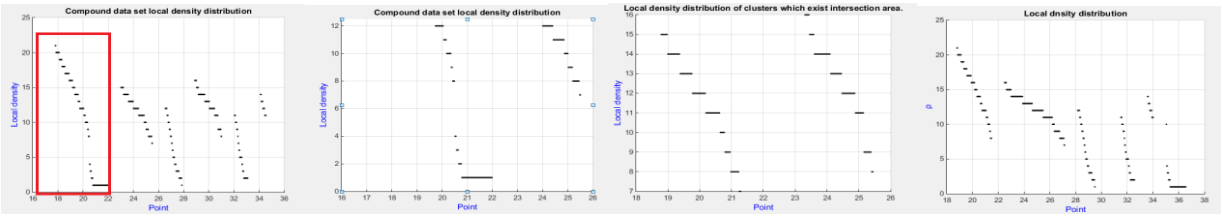
(a) Data set distribution (b) Decision graph (c) Clustering result (d) Cluster centers

Fig.2: The CFSFDP algorithm is examples on the Compound data set

We can clearly find that the algorithm incorrectly clustering the data sets in Fig.2(c)(marked by the frames). Fig.2 (d) show the location of the cluster centers in the data set, it can be seen that the two clusters with different densities share the same density peak(marked by red frame in Fig.2(c)) and the annular cluster have two density peaks(marked by blue frame in Fig.2(c)). So we need find a approach to solve the problems.

2.2. Division clusters

In the previous section we introduced and analyzed the CFSFDP algorithm. The reason for the incorrect clustering in Fig.2(c)(marked by red frame) is because the point in the low density region is attached to the high density region and can't get high value of δ, so it is not possible to find the cluster center of the low density region from the decision graph. We propose a approach to describe the clustering structure by describing the density distribution of the clusters similar to the OPTICS algorithm. Sorted the points of the clusters in descending order according to the local density, the clustering structure of the data set can be clearly observed. Fig.3 (a) shows the Compound data set local density distribution, it can be clearly found that the local density distribution for cluster 0(marked by red frame in Fig.3 (a)) is divided into two segments, this means that the cluster is composed of two clusters with large differences in density, enlarge the density distribution of cluster 0 and displayed in Fig.3 (b), According to the graph, we can select the appropriate density to division the cluster.



(a) Density distribution (b) Enlarged view of Fig3.(a) (c) Intersection clusters (d) Density distribution

Fig.3: Compound data set local density distribution

2.3. Merge clusters

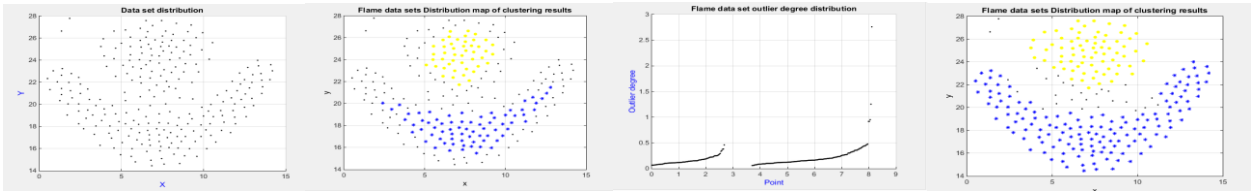
For a cluster containing multiple density peaks, CFSFDP algorithms will excessive division the cluster. We use a hierarchical clustering method based on greedy thought to merge the clusters, Fig3.(c) shows the local density distribution of clusters of the Compound data set which exist intersection region. it can be seen that these two clusters have a similar density distribution, if the average density of the intersection region can make the two clusters satisfy the density link, we consider there are necessary to be merged, and the method can make the number of clusters as small as possible. Fig3.(d) shows the local density distribution after division and merge operation, we can see that each cluster has a continuous and evenly density distribution. and the data set is correctly clustered into six clusters. The merger judgment condition defined as Eq (2-4).

$$\rho_{avg}(c_i) = \frac{1}{|\,c_i\,|}\sum_{x\in c_i}\rho_x$$
$$B_{c_i,c_j} = \{x \mid x \in c_i, \exists z \in c_j, d_{xz} < d_c\} \cup \{y \mid y \in c_j, \exists z \in c_i, d_{yz} < d_c\} \quad (2\text{-}4)$$
$$\rho_{bound}(c_i,c_j) = \frac{1}{|\,B_{c_i,c_j}\,|}\sum_{x\in B_{c_i,c_j},}\rho_x$$

2.4. Outlier degree

According the CFSFDP algorithm, for the noises, we need find the point of highest density within its border region for each cluster. We denote its density by ρ_b, the points of the cluster whose density is lower than it are considered as noise. Fig.4 (a) shows the result of Flame [11] data set clustering by CFSFDP algorithm (the noises marked by bleak point), we find that there are too many points determined as noise, this is because the calculated ρ_b is usually large. By observing the decision graph, we found the noise generally has a low ρ and a high δ simultaneously, which deviates from the distribution of other points in the cluster [12]. We denote the outlier degree by OD, it is defined as Eq.(2-5).

$$OD_i = \frac{\delta_i}{\rho_i} \quad (2\text{-}5)$$



(a) Flame data set (b) CFSFDP (c) Outlier degree distribution (d) DMCFSFDP

Fig.3: Compound data set local density distribution

In Fig3. (c), we describe the outlier degree of the points in the data set. It can be obviously found that some points deviates from the distribution of other points and has a high OD, we think these points can be considered as noise. Fig3. Select an appropriate outlier degree as the threshold, and delete the points above that value. Fig3. (d) shows the clustering result by using outlier degree to determine noise, it can be seen that it better maintains the clustering structure and avoids large amounts of data to be deleted.

2.5. DMCFSFDP

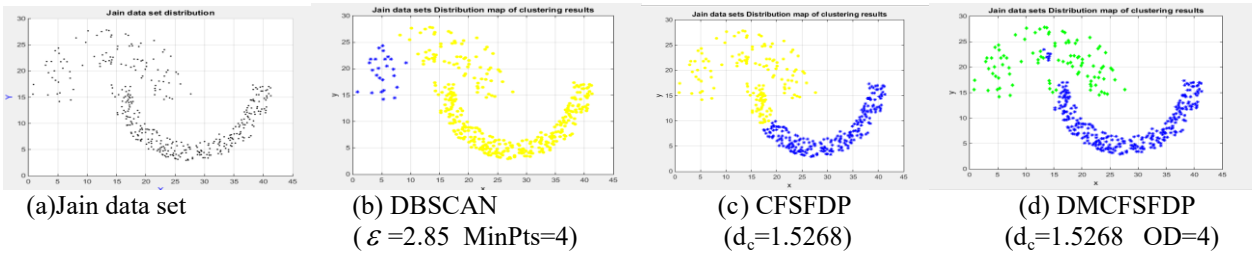
Through the above analysis, we build the DMCFSFDP algorithm as the following steps .

- Calculate the ρ and δ values for each data point.
- Build decision graphs and select the appropriate number of cluster centers.
- The rest points are assigned to the same cluster as its nearest neighbor of higher density.
- Construction the local density distribution of each cluster, division the clusters with large differences in density.The divided clusters are no longer merged with the original cluster.
- Merge the clusters which exist intersecting regions by Eq(2-4).
- Determined noise according outlier degree of the point.

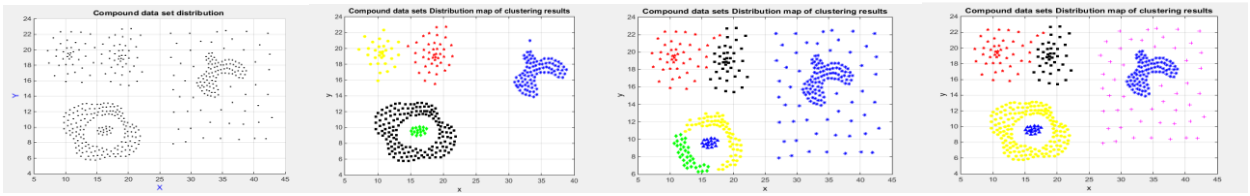
3. Experiment and analysis

3.1. Experiment

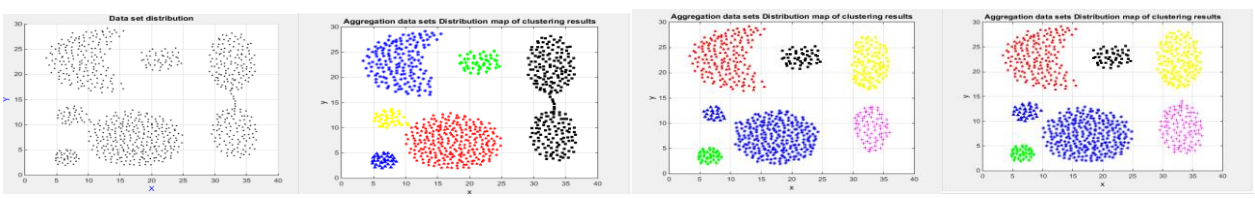
We will verify the validity of the DMCFSFDP algorithm by comparing with DBSCAN algorithm and the CFSFDP algorithm. Using the recognized data set as the algorithm input. The clustering results are shown in Fig.5. Parameters are given below the graphs and noises are not displayed in the results.



(a)Jain data set (b) DBSCAN (ε =2.85 MinPts=4) (c) CFSFDP (d_c=1.5268) (d) DMCFSFDP (d_c=1.5268 OD=4)



(a)Compound data set (b) DBSCAN (ε =1.20 MinPts=4) (c)CFSFDP (d_c=1.4671) (d) DMCFSFDP (d_c=1.4671 OD=2)



(a)Aggregation data set (b) DBSCAN (ε =1.20 MinPts=4) (c)CFSFDP (d_c=1.8594) (d) DMCFSFDP (d_c=1.8594 OD=0.6)

Fig.4: Clustering results of data sets under different algorithms

3.2. Algorithm analysis

Limited by the length of this paper, we selected part of the experiment results to show. The Jain [13] data set have two clusters of different densities and shapes, and the Compound data set have six clusters with different densities and shapes. Through a large number of experiments, we found that DBSCAN algorithm is very sensitive to the parameters, and can not identify clusters with large differences in density effectively. For the CFSFDP algorithm, the algorithm performances well on multiple data sets and for the threshold d_c is robust, we found the algorithm is very accurate to identify clusters with single density peaks. But as the Fig.4(c) shown that the algorithm cannot effectively identify clusters with low density or multiple density peaks, and determines too many data points as noise. Fig.4(d) shows the clustering results calculated by the DMCFSFDP algorithm, It can be found that the algorithm can effectively find clusters of arbitrary shape and density, and identify the noises more accurately.

4. Conclusions

In this paper, we proposed a approach to improve the CFSFDP algorithm, the algorithm based on the density distribution and outlier degree. Experiments show that the algorithm is better than CFSFDP in identifying clusters with different density and arbitrary shapes, and the algorithm has the same time complexity with the CFSFDP algorithm. Although the algorithm solves some disadvantages of CFSFDP algorithm, but still have some research area, Such as optimizing the threshold d_c or automatically select the initial cluster centers and so on. So, we will start research in these areas in the future.

5. References

[1] B Li, D Jiang. The Research of Intrusion Detection Model Based on Clustering Analysis [C]//*International Conference on Computer and Communications Security,2009：24-27.*

[2] PARDESHI B， TOSHNIWAL D．Improved K-medoids clustering based on cluster validity index and object density［C］//*Proc of the 2nd IEEE International Advance Computing Conference．2010: 379-384．*

[3] G.Karypis,E.-H Han,and V.Kumar.Chamlenon:A hierarchical clustering algorithm using dynamic modeling.*COMPUTER,1999(32):68~75*

[4] Sudipto Guha,Rajeev Rastogi,and Kyuseok Shim.Cure:An efficient clustering algorithm for large database.*In:Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,1998.73~84*

[5] ESTER M , K RIEGEL H , SANDER J, et al.A density-based algorithm for discovering clusters in large spatial databases with noise[C] // *Proc of the 1996 2nd Int'l Conf on Knowledge Discovery and Data Mining .Portland :AAAI Press , 1996:226-231.*

[6] M.Ankerst,M.Breunig,H.-PKriegel,and J.Sander.Optics:Ordering points to identify clustering structure.*in:Proceedings of the ACM SIGMOD Conference,pages 49-60,Philadelphia,PA,1999.102~123*

[7] G.Sheikholeslami,S.Chatterjee,and A.Zhang.Wavecluster:A multiresolution clustering approach for very large spatial databases.*in: Proceedings of the 24th Conference on VLDB,New York,NY,1998.428~439*

[8] A. Rodriguez and A. Laio,"Clustering by fast search and find of density peaks,"*Science,vol.344,pp.1492-1496,June.2014.*

[9] Gionis,A.,H.Mannila,and P.Tsaparas,Clustering aggregation.ACM Transactions on Knowledge Discovery from Data(TKDD),2007.1(1):p.1-30.

[10] C.T. Zahn, Graph-theoretical methods for detecting and describing gestalt clusters.*IEEE Transactions on Computers, 1971. 100(1): p. 68-86.*

[11] Fu,L.and E.Medico,FLAME,a novel fuzzy clustering method for the analysis of DNA microarray data.*BMC bioinformatics,2007.8(1):p.3.*

[12] M.M.Breunig,H.P.Kriegel,R.T.Ng,J.Sander.LOF:Identifying Density-Based Local Outliers.*In Proceedings of the 2000 ACM SIGMOD international Conference on Management of data.2000.29.93-104*

[13] Jain,A.and M.law,Data clustering:A user's dilemma.*Lecture Notes in Computer Science,2005.3776:p.1-10*