Density Peaks Clustering for Complex Datasets

Shanshan Ruan
College of Information Science and Technology
Beijing Normal University, Beijing 100875, China

Saeed El-Ashram'*
State Key Laboratory for Agrobiotechnology and College of Veterinary Medicine,
China Agricultural University, Beijing 100193, China
,Faculty of Science, Kafr El-Sheikh University,
Kafr El-Sheikh, Egypt Email:saeed_elashram@yahoo.com

Waqas Ahmad
College of Information Science and Technology
Beijing Normal University, Beijing 100875, China

Zahid Mahmood
School of Computer and Communication Engineering,
University of Science and Technology Beijing (USTB), Beijing, 10008, China

Rashid Mehmood
College of Information Science and Technology
Beijing Normal University, Beijing 100875, China

Abstract—**Clustering by fast search and find of density peaks (DP) is a new density based clustering method and has gained much popularity among the researcher. DP provided the new insight to detect cluster centers and noise in the dataset. DP reveals that a cluster center is a point that have higher density as compared with its neighbor points and have a large distance from other higher density peak points. DP detects each density peak in dataset and discover cluster center with the help of decision graph with minimum human interpretation. After successful identification of cluster centers, rest of points are assigned to each cluster center based on the nearest neighbor of higher density. DP works very well when each cluster consists of single density however, for more complex and density connected clusters, it cannot finds the accurate clusters. To make DP effective equally for more complex datasets, we introduce a novel approach to detect miss classified density and then assign separate density to appropriate cluster. To evaluate the robustness of proposed method, we utilized three complex synthetic datasets and compared with DP.**

Keywords-clustering;density peaks;complex datasets

I. INTRODUCTION

Clustering analysis is a method to organize similar data into different groups and has become an important research field of the data mining[1]. It has been successfully applied in various area, such as bioinformatics, environment, marketing, education, machine learning, and pattern recognition. Many clustering algorithms have been proposed and studied in the literature however, the choice of appropriate cluster method is highly depends upon the nature of underlying dataset. Mostly, clustering algorithms need many input parametric setting to create effective clusters, the shape and number of resultant clusters are based on the input parameters. However, an ideal clustering method should be effective, robust, and capable to organize the data into appropriate groups, and take minimum input parameters.

K-means is one of the most simple and state of the art partitioning based clustering algorithm, it has applied and studied in many fields. It partition the whole dataset into predefined number of clusters and iteratively refines each partition to get meaningful clusters. Its key drawbacks includes:(1) hard to assess how many clusters exist in data; (2) selection of initial cluster centers; (3) cannot finds arbitrary shape of clusters; (4) not sensitive for noise or outliers.

DBSCAN is an effective density based clustering algorithm to find arbitrary shape of clusters with the application of noise or outliers. DBSCAN takes two input parameters, min-points and min-distance, to find the arbitrary shape of clusters. The min-points is how many points should be inside a min-distance radius from other points to be consider as a part of the existing cluster. However, DBSCAN stuck to cluster overlapping densities and drops the density at border region of cluster[2].

Recently, clustering by fast search and find of density peaks (DP) has proposed by Rodriquez et al. [3], which have the characteristics of partition and density based methods. DP partition the dataset around the centroid and then assign rest of points to cluster center based on the minimum distance to nearest cluster center. Firstly, DP estimates the

978-1-5090-5952-2/16 $31.00 © 2016 IEEE
DOI 10.1109/IIKI.2016.20

local density and then it calculates the minimum distance of all points from the nearest maximum density. In this way, the cluster centers are identified as points, have higher density and higher distance . DP utilizes an approach named as decision graph to detect cluster centers. On decision graph the only cluster centers have higher distance and density, hence became the outliers as compared with rest of points, outliers easily can be identified as cluster centers. DP is sensitive for relative density of data, the shape and number of clusters are depends upon the estimation of density[4]. Many invariant of DP has been proposed to simplify the selection of input parameters. In CFSFDP-HD [4], an adaptive approach has been proposed to estimate the density adaptively using heat-diffusion method and dependency of cutoff distance (d_c) has solved successfully. In adaptive fuzzy-CFSFDP [7] has proposed to find the exact number of clusters, adaptively. In FKNN-DPC [5]is based on the nearest neighbor approach, a high dense point have maximum neighbors and points are assigned to cluster centers based on the nearest neighbor based approach. Some other approaches such as [6, 8, 9] also have proposed to overcome the limitation of DP.

The rest of this paper is organized as follows. The related work is presented in Section 2. Section 3 describes the proposed DP-FCD in detail. Experimental results and comparisons are given and discussed in Section 4, and finally, the concluding remarks and future work is presented in Section 5.

II. Related Work

Clustering by fast search and find of density peaks (DP) has proposed to detect cluster centers, exclude outliers, and organized data into different clusters regardless of their shape and dimensions. DP is primarily based on the following two assumptions that the ideal cluster centers: (1) have high density as comparing with their surrounding neighbors, and (2) have high distance from any point with a high local density. To detect the cluster center, DP calculates the local density (ρ_i) and distance(δ_i) of each point i . The $definition - 1$ is utilized to estimate the local density, which is given as follows: $Definition - 1$:

$$\rho_i = \sum_j X\left(d_{ij} - d_c\right), \tag{1}$$

where

$$X(d) = \begin{cases} 1 & d < 0 \\ 0 & otherwise. \end{cases}$$

Where d_c is the cutoff distance d_{ij} is distance of point point i to j . Simply, ρ_i is equal to number of points that are more close then d_c to point i . The distance (δ_i) of each data point i can be calculated using the $definition - 2$, given as follows: $Definition - 2$:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} \left(d_{ij}\right) & if \ \exists \ j \ s.t. \rho_j > \rho_i \\ \\ \max\limits_{j:\rho_j > \rho_i} \left(d_{ij}\right) & otherwise. \end{cases} \tag{2}$$

After calculation of ρ_i and δ_i for each point i , DP plots the ρ_i and δ_i to discover the cluster centers, named as decision graph. DP suggests that the outlier at decision graph can be selected as cluster centers. After successful selection of cluster centers, rest of points are assigned to nearest cluster centers based on the minimum distance from point i to cluster center C_i, and then noise is removed from each cluster. DP works well on clusters consisted of single density, however in complex datasets, it misclassify data and assign whole connected density to nearest cluster center without finding the density connection between two connected densities, as shown in fig-1(b), fig-2(b),and fig-3(b).

To overcome the aforementioned limitation of DP, we proposed an novel approach named as Density Clustering for complex datasets (DP-FCD).

III. Proposed Method

The DP-FCD is proposed to overcome the limitation of DP to accurately classify the complex datasets. As shown in fig. 1(b), DP partition the dataset into different partitions and then assigned each partitioned to cluster center based on the nearest distance, and ignore the connected densities. Specially, in red cluster the outer ring partition is belonged to orange color cluster but DP misclassified and merged it into red cluster because of nearest distance from cluster center. The proposed method is sensitive to detect the connectivity among the connected densities. The DP-FCD works in three steps that are: (1) to detect separated density; (2) check the density connection; and (3) refine the overlapping region.

Step-1 detect separated density: DP assigne the points to cluster centers based on the nearest distance from cluster centers but ignore the connected densities. So for complex datasets it mislead to assign the density to appropriate cluster center. To handle this, we firstly detect the connected densities in each cluster. We utilize the following algorithm-1 to mark the connected densities to a cluster.

Algorithm 1 : Detect the nearest neighbors that are far away from d_c distance.

```
Require:  $d_c$, the cutoff distance
$D$, all pair distance matrix
Ensure:  $\rho$, the density vector of all points $i$
    $for\ i \leftarrow 2{:}n\ ,do$
    $\delta(ordraho(i)) \leftarrow \max(D)$
    $for\ j \leftarrow 1{:}i-1\ ,do$
        $if\ dist\big(ordrho(i), ordrho(j)\big) < \delta\big(ordrho(i)\big), then$
            $\delta(ordraho(i)) \leftarrow dist\big(ordrho(i), ordrho(j)\big)$
            $nneigh(ordrho(i)) \leftarrow ordrho(j)$
            $if\ dist\big(ordrho(i), ordrho(j)\big) < d_c, then$
                $Tempnneigh(ordrho(i)) \leftarrow ordrho(i)$
            Endif
        Else
            $\delta_i \leftarrow \max(dist(i,j)$
        $Endelse$
        Endif
        Endfor
        Endfor
```

$Tempnneigh$ contains all the data points that are assigned to single cluster center based on the nearest neighbor distance without checking the connected densities. Hence, it might be possible that these points belong to other cluster. As shown in fig.2(b), outer spiral is connected to red cluster, however, actually it belongs to pink cluster. In our proposed work, $C_{i\_Tempnneigh}$ contains all those paints that have their index in $Tempnneigh$ and might be misclassified by DP.

Step-2: Discover density connections: To discover the connected densities, we check how many points are at a d_c distance from other cluster border region and assign each separate density $C_{i\_Tempnneigh}$ to cluster have more nearest neighbors within d_c distance at border region.

Step-3: Refine the assigned points: After assignment of $C_{i\_Tempnneigh}$ to appropriate cluster, we again check the probability of overlapping points and refine them according to their density to nearest cluster.

IV. Experiments

To evaluate the performance of our proposed method, we used Compound[10], Jain[11], and Path based [12] synthetic datasets. The compound dataset consisted of five clusters and have 399 two dimensional data points, as shown in fig.1 (a). At optimized selection of d_c and number of clusters, DP could not detect the connected densities of red and orange clusters and merge a large part of orange cluster into red cluster, as shown in fig.1(b). This is happened because DP checks only the nearest density but ignore the high probability of connectivity that exists between densities. So in our proposed work, we first identify the misclassified points and then check the degree of connectivity of disputed density. We marked as disputed region that have might be part of other cluster, as shown in fig.1(c). Based on the maximum connectivity we assigned the disputed density to orange cluster. However, to handle the disperse points we assigned each point to nearest neighbor at border region based on the similarity of densities and minimum distance. We take each disputed region and then assigned to nearest cluster based on the maximum connectivity at border region. For sparse points, we test the probability based on the nearest neighbor and density then assigned to cluster having maximum similarity to that point, as shown in fig.1(d).

The path-based dataset consists of three clusters and have 300 two dimensional data points, as shown in fig.2 (1). DP successfully identify exact cluster centers but while assigning remaining points to nearest centers it could not organize the whole datasets into appropriate clusters , effectively. DP assigned independent densities based on the nearest distance and ignored the connectivity, as shown in fig.2(2). In fig.2(b), at both side the spiral structure belongs to pink cluster, however DP assigned to red and yellow clusters because of nearest distance. In proposed work, we first detect all points that have probability to misclassification. In fig.2 (3), black color points show the densities that have a probability of miss classification. In next step, we checked the connectivity with other clusters and assigned based on the nearest neighbor. Figure 2(d) shows the final organized clusters of proposed method.

To benchmark the proposed method on more complex datasets, we utilized the Jain dataset that consisted of two overlapping clusters having 373 data points, as shown in fig.3(a). DP created clusters are shown in fig.2(b), where a portion of green cluster is merged into red cluster. Unlike DP, the proposed method firstly discover all connected densities and then assign each density to cluster based on the nearest neighbor and then refine the overlapping points that lies in both clusters. The final organized cluster of proposed method are shown in fig.3(d).

V. Conclusions

DP clustering has been proposed to identify automatic clusters, however it works on simply datasets and could not cluster complex datasets exactly. To overcome the assignment of remaining points to cluster centers, we proposed a new method to exactly assign remaining points to cluster centers. Our approach successfully identify the different densities that have a probability to misclassification. After identification of independent densities, we check the connectivity among clusters and assign one by one each density to nearest cluster. The tested results on three complex datasets validate the robust and effectiveness of proposed method.

In future we will plan to introduce the automatic estimation of densities and detect cluster centers without depending upon the decision graph approach.

Acknowledgment

This research is sponsored by National Natural Science Foundation of China (No.61171014,61371185, 61401029,61472044,61472403,61571049) and the Fundamental Research Funds for the Central Universities (No.2014KJJCB32, 2013NT57) and by SRF for ROCS, SEM.

References

[1] Jiao L, Zhang G, Wang S, Mehmood R, Bie R. Optimal Preference Detection Based on Golden Section and Genetic Algorithm for Affinity Propagation Clustering. InInternational Conference on Wireless Algorithms, Systems, and Applications 2015 Aug 10 (pp. 253-262). Springer International Publishing.

[2] Campello, R. J., Moulavi, D., and Sander, J. (2013, April). Density-based clustering based on hierarchical density estimates. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 160-172). Springer Berlin Heidelberg.

[3] Rodriguez, Alex, and Alessandro Laio,Clustering by fast search and find of density peaks, Science,vol.344, no. 6191,pp: 1492-1496,2014.

[4] Mehmood R, Zhang G, Bie R, Dawood H, Ahmad H, Clustering by fast search and find of density peaks via heat diffusion, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.01.102i

[5] Xie, Juanying, Hongchao Gao, Weixin Xie, Xiaohui Liu, and Philip W. Grant. "Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors." Information Sciences 354 (2016): 19-40.

[6] Bie R, Mehmood R, Ruan S, Sun Y, Dawood H. Adaptive fuzzy clustering by fast search and find of density peaks. Personal and Ubiquitous Computing. 2016 Oct 1;20(5):785-93.

[7] Mehmood R, Bie R, Jiao L, Dawood H, Sun Y. Adaptive cutoff distance: Clustering by fast search and find of density peaks. Journal of Intelligent and Fuzzy Systems. 2016 Jan 1;31(5):2619-28.

[8] Ruan Shanshan, Rashid Mehmood, Jalal Alowibdi, Zhang Zhongjun, Hussain Dawood and Ali Daud, An adaptive method for clustering by fast search-and-find of density peaks, In proceeding of world wide web conference, March, 2017.

[9] Mehmood R, Bie R, Dawood H, Ahmad H. Fuzzy Clustering by Fast Search and Find of Density Peaks. In2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI) 2015 Oct 22 (pp. 258-261). IEEE.

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

[11] Jain and M. Law, Data clustering: A user's dilemma. Lecture Notes in Computer Science, 2005. 3776: p. 1-10.

[12] H. Chang and D.Y. Yeung, Robust path-based spectral clustering. Pattern Recognition, 2008. 41(1): p. 191-203.

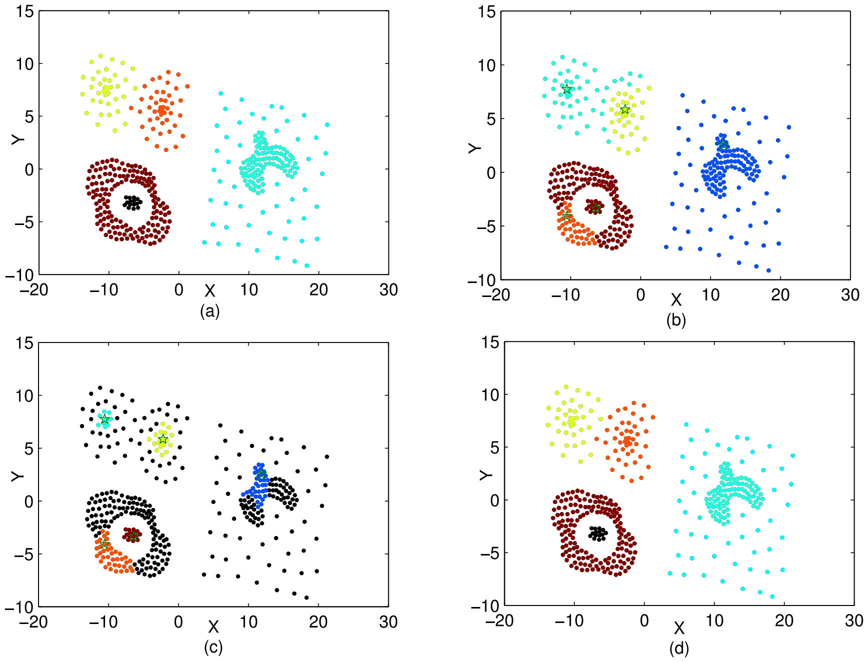


Figure 1: Different stages of proposed method on compound dataset and comparison with DP. (a) Shows the ground truth of compound dataset, (b) Shows the five organized clusters of DP ,at d_c=1.(c) Illustrate the identification of disputed density regions of clusters.(d) Shows the organized clusters of proposed method, at d_c=1

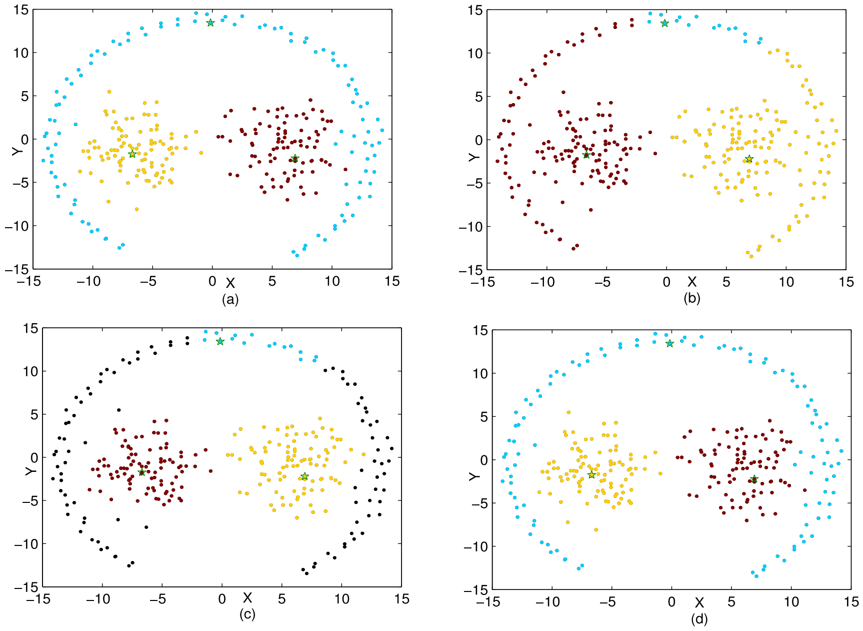
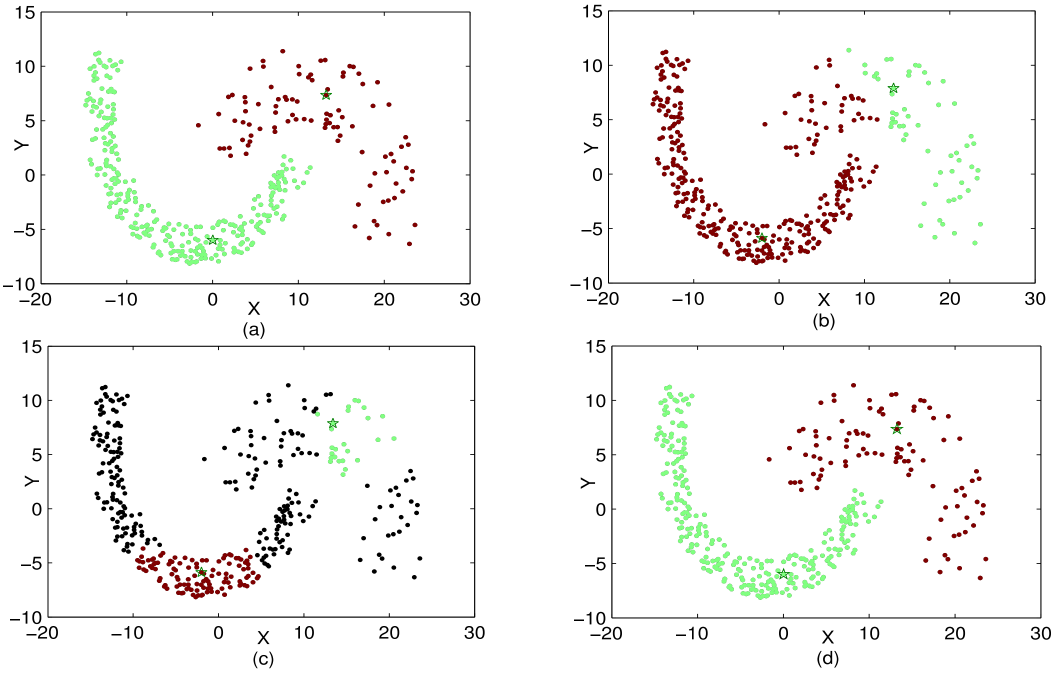


Figure 2: Comparison of proposed method and DP on Path-based dataset. Path-based dataset having 300 two dimensional data points, organized into three clusters, as shown in (a). (b) Shows the organized clusters of DP. (c) Shows the misclassified points of DP that are marked as black color. (d) Organized clusters of proposed method.



(a)

(b)

(c)

(d)

Figure 3 : DP and proposed method comparison on Jain data. (a) Shows the facts of Jain dataset. (b) Shows the DP clusters of Jain dataset that could no classify exactly. (c) Shows the detected of points that might a disputed region. (d) Shows the final clusters of organized by proposed method.