Enhancing Cluster Center Identification in Density Peak Clustering

Jian Hou[1(✉)], Aihua Zhang[1], Chengcong Lv[1], and Xu E[2,3]

[1] College of Engineering, Bohai University, Jinzhou 121013, China
dr.houjian@gmail.com
[2] College of Information Science, Bohai University, Jinzhou 121013, China
[3] College of Food Science and Technology, Bohai University, Jinzhou 121013, China

Abstract. As a clustering approach with significant potential, the density peak (DP) clustering algorithm is shown to be adapted to different types of datasets. This algorithm is developed on the basis of a few simple assumptions. While being simple, this algorithm performs well in many experiments. However, we find that local density is not very informative in identifying cluster centers and may be one reason for the influence of density parameter on clustering results. For the purpose of solving this problem and improving the DP algorithm, we study the cluster center identification process of the DP algorithm and find that what distinguishes cluster centers from non-density-peak data is not the great local density, but the role of density peaks. We then propose to describe the role of density peaks based on the local density of subordinates and present a better alternative to the local density criterion. Experiments show that the new criterion is helpful in isolating cluster centers from the other data. By combining this criterion with a new average distance based density kernel, our algorithm performs better than some other commonly used algorithms in experiments on various datasets.

Keywords: Clustering · Density peak · Local density · Cluster center

1 Introduction

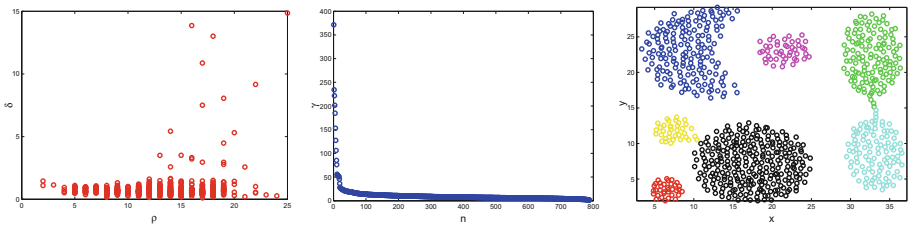
Data clustering has wide applications in such fields as data mining, pattern recognition and others. Many clustering algorithms of different types have been developed and some of them have generated impressive results in application. Some commonly used algorithms include k-means, spectral clustering [13,16], DBSCAN [7], mean shift [5] and their variants. Recently, some new algorithms have been proposed, including affinity propagation (AP) [3], robust spectral clustering [19], dominant sets (DSets) [14]. Noticing that many algorithms require to determine the number of clusters beforehand, [8,12] have presented some methods to solve this problem. Since some algorithms detect only spherical clusters, density based algorithms have received a lot of attention [1,2].

W. Liu et al. (Eds.): KSEM 2018, LNAI 11061, pp. 268–275, 2018.
https://doi.org/10.1007/978-3-319-99365-2_24

In density based clustering algorithms, DBSCAN relies on a density threshold to detect cluster borders, and the density threshold is represented by two parameters $MinPts$ and Eps. While DBSCAN has been shown to perform well in many experiments, it may not be easy to determine the appropriate parameters. In addition, a fixed set of parameters imply a fixed density threshold, which may not be appropriate for datasets where cluster densities vary significantly. Different from DBSCAN-like algorithms, the density peak (DP) algorithm presented in [15] accomplishes the clustering process on the basis of density relationship. By treating local density peaks as the candidates of cluster centers, the DP algorithm finds that cluster centers have both great ρ's and great δ's, and either the ρ's or δ's of non-density-peak data are small. This algorithm then uses both ρ and δ, or $\gamma = \rho\delta$, to identify cluster centers, and then group the other data into clusters based on density relationship among neighboring data. Different from cluster centers surrounded by data of smaller density, non-density-peak data usually have the nearest neighbors greater density in the neighborhood, corresponding to small δ's. Consequently, the distance δ is effective in isolating cluster centers from the non-density-peak data. While cluster centers usually have greater local density than the neighboring non-density-peak data, the density of non-density-peak data may not be small in absolute magnitude. In other words, non-density-peak data may have great or small local density, and the ρ criterion is not very informative in strengthening the specificity of cluster centers. For the purpose of solving this problem, we study the cluster center identification process and find that the role of density peak is more important for a cluster center than a great density. We then present an enhanced criterion based on local density of subordinates to replace the original local density ρ. Furthermore, a new density kernel is proposed to overcome the drawbacks of the cutoff and Gaussian kernels. By combining the new criterion and density kernel, our algorithm performs well in experiments and compares favorably to some commonly used and recently proposed algorithms.

2 Density Peak Clustering Algorithm

An attractive property of density based clustering algorithms is that they detect non-spherical clusters. While the DBSCAN algorithm based on a density threshold, the DP algorithm makes use of the density information in a different manner. We use examples to demonstrate how the DP algorithm identify cluster centers and accomplish the clustering process. With the Aggregation [10] dataset, we calculate ρ in the first step. The cutoff kernel defines the local density as the count of data in the d_c-radius neighborhood, where d_c is the cutoff distance to be specified. The distribution of ρ and δ of all the data is shown in Fig. 1(a). Obviously only a few data have both great ρ and great δ and the majority of the data have either small δ's or small ρ's. This makes it feasible to determine cluster centers by selecting the data of great ρ's and great δ's. Noticing that with Fig. 1(a) two thresholds are necessary to determine cluster centers, we sort the data according to $\gamma = \rho\delta$ in the decreasing order and show the distribution of γ



(a) ρ-δ decision graph (b) γ decision graph (c) Clustering result

Fig. 1. Decision graphs and clustering results with the cutoff kernel.

in Fig. 1(b), where a few data are with significantly greater γ than the others. With the γ decision graph we only need one threshold to select out the cluster centers.

With the cluster centers available, the DP algorithm determines the labels of the non-density-peak data based on data density relationship. The clustering of non-density-peak data is on the basis of the assumption that one data and the nearest neighbor of greater local density are in the same cluster. With this assumption, the non-density-peak data can be assigned labels in the decreasing order of their local density. This process involves only one scan of the data and can be accomplished efficiently. While no proof shows that this assumption holds in theory, the method works well in practice, as shown in the clustering results in Fig. 1(c).

While Fig. 1 shows that cluster centers usually have great ρ, great δ and great γ, they also indicate that it may not be easy to select out the cluster centers with only the decision graphs. The reason is that the differences between great and small ρ's, δ's and γ's are not significant in many cases. For the purpose of avoiding the influence from inappropriate thresholds, we specify the number of clusters in this paper, and the data of the greatest γ's are identified as cluster centers.

3 Our Approach

Since cluster centers typically have both great ρ's and great δ's, we often uses $\gamma = \rho\delta$ to identify cluster centers. As density peaks, cluster centers are surrounded by neighboring data of smaller local density. As a result, they are distant from the nearest data of greater local density and have great δ's. However, the nearest neighbors of cluster centers may have only slightly smaller density than the correspondingly cluster centers. In other words, non-density-peak data may also have great density. As Fig. 1(a) shows, the local densities of the data are distributed quite evenly, and there are a large amount of data with great local density. This observation indicates that the local density ρ is not as informative as δ in isolating cluster centers from non-density-peak data. In our opinion, this also explains why the DP clustering results are influenced by density kernel types and kernel parameters significantly.

In order to relieve the problems resulted by the uninformative ρ criterion in identifying cluster centers, we make a further study of the cluster center identification process. One intention of the DP algorithm is to use some measures to strengthen the specificity of cluster centers. Since cluster centers have both great ρ's and great δ's, and either the ρ's or δ's of non-density-peak data are small, the product $\gamma = \rho\delta$ is used as the cluster center identification criterion. We have observed that δ is indeed effective for cluster center identification, and ρ is not so informative in comparison. However, if we remove ρ and uses only δ to select cluster centers, it is likely that the outliers of datasets which are far from other data are identified as cluster centers. In other words, the local density ρ is still effective in preventing outliers from being identified as cluster centers. Therefore instead of removing ρ completely, we propose to enhance the discriminative ability of ρ.

In the DP algorithm, one important feature of cluster centers is that they are surrounded by non-density-peak data of smaller local density. Here we see that what differentiates cluster centers from non-density-peak data is not the great density in absolute magnitude, but the role of density peaks. That is to say, it doesn't matter if one cluster center has a great density, but it matters if it has a greater local density than the neighboring data. Hence we propose to use a criterion measuring the role of density peaks to replace ρ in identifying cluster centers.

In the following we take one data i for example, and denote the cluster containing i as C_i. If i is the cluster center of C_i, it should be surrounded by neighboring data of smaller density. Intuitively we can use the number N_n of neighboring data with smaller density to measure the role of i being the cluster center. A larger N_n means a larger possibility of i being the cluster center of C_i. However, it is possible that the neighboring data with smaller density contain not only the data in C_i, but also some data in other clusters. In this case, N_n cannot measure the possibility of i being the cluster center accurately, and we need to consider only the data in C_i. However, before the clustering is accomplished, the cluster membership of C_i is unknown.

We present the following method to make use of only the data in C_i before the cluster membership is available. It is assumed in the DP algorithm that one data and the nearest data of greater local density are in the same cluster. If one data n is the nearest neighbor of greater local density of data m, we call n as the *superior* of m, and m as the *subordinate* of n, and denote this relationship by $m \rightarrow n$. Evidently one data and all its subordinates should be in the same cluster. Since cluster centers are density peaks, they usually have a large amount of subordinates. On the contrary, the number of subordinates of non-density-peak data may be small or zero. Therefore for the data i, we can use the number N_s of subordinates to measure the probability of i being the cluster center. Furthermore, the local density of subordinates also plays a role in measuring the possibility. In summary, we use the sum of local density of the subordinates to measure the possibility of i being the cluster center, and define the enhanced version of ρ as

$$\eta_i = \sum_{j \in S, j \neq i} \zeta(i,j)\rho_j, \tag{1}$$

where

$$\zeta(x,y) = \begin{cases} 1, & y \to x, \\ 0, & otherwise. \end{cases} \tag{2}$$

Then we can use η to replace ρ and identify cluster centers based on $\gamma' = \eta\delta$. It is worth mentioning that we only use η in determining the cluster centers. The original ρ is still used in grouping non-density-peak data on the basis of density relationship, as it measures the density relationship among neighboring data more accurately.

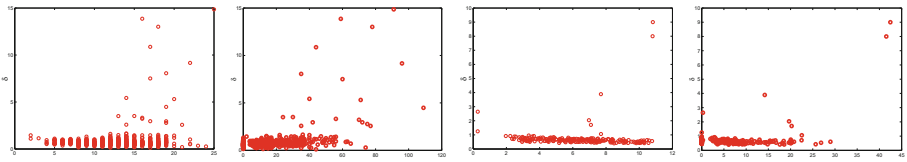
In addition, we make use of the average distance to a limited amount of nearest data to evaluate the local density. The density kernel obtained this way is presented as a compromise between the cutoff and Gaussian kernels. The cutoff kernel makes use of only the count of data in a neighborhood and discards the distance information to these data. This information loss may influence the local density precision. While the Gaussian kernel makes use of the distance information, it takes into account both the nearest neighbors and the farthest data. In this case, the density kernel may measure the distribution of data in a large region but not a small neighborhood, if the parameter d_c is not selected appropriately. Between these two extremes, our new kernel makes use of the distance to a limited number of neighboring data, and is shown to perform well in experiments.

4 Experiments

In our work η is presented as an enhanced version of local density ρ to improve the discriminative ability, and then use a new density kernel to overcome the drawbacks of existing ones. In this part we firstly validate the effectiveness of the enhanced local density criterion. After that, the whole algorithm is tested and compared with existing commonly used and recently proposed algorithms.

4.1 Enhanced Local Density

The ρ-δ decision graph in Fig. 1 shows that the distribution of data in the range of the local density ρ is quite even, indicting that ρ is not very informative in strengthening the specificity of cluster centers. We are motivated to replace ρ by η to help isolate cluster centers from non-density-peak data. Here we test if η really works in serving this purpose. By replacing ρ with η, we show the η-δ decision graphs and the corresponding ρ-δ decision graphs on the Aggregation and Flame datasets in Fig. 2. Evidently the majority of the data have small η values, and only a few data are with great η. The comparison between ρ-δ graphs and η-δ graphs indicates that η is helpful in isolating cluster centers from non-density-peak data, and is more suitable to serve as a cluster center identification criterion than ρ.



(a) ρ-δ (b) η-δ (c) ρ-δ (d) η-δ

Fig. 2. The η-δ decision graphs and corresponding ρ-δ decision graphs. The left two figures belong to the Aggregation dataset, and the right two correspond to the Flame dataset.

4.2 Comparison

We now compare the proposed algorithm with some commonly used and recently proposed clustering algorithms with experiments on eight datasets, including Aggregation, Compound [18], Spiral [4], D31 [17], R15 [17], Flame [9] as well as the Wine and Iris datasets from UCI machine learning repository. Aside from the well-known k-means and DBSCAN algorithms, the normalized cuts (NCuts) [16], AP, DSets, one improved version of DSets presented in [11] and SPRG [19] are also adopted in comparison. Since our work is proposed to improve the DP algorithm, we also compare with two versions of the DP algorithm, one of which with cutoff kernel (DP-c) and the other with Gaussian kernel (DP-G). We experiment on the same eight datasets as in previous sections, and report clustering results evaluated with NMI. Except for the algorithm in [11], all these algorithms require to input one or more parameters. The k-means, SPRG and NCuts algorithms involve the number of clusters, and we set this parameter as the ground truth. As to DBSCAN which has two parameters $MinPts$ and Eps, we set $MinPts = 3$ which is selected from 2, 3, $\cdots$, 10, and then determine Eps based on $MinPts$ [6]. The AP algorithm involves the preference value p, and the authors of [3] provide a method to obtain the range $[p_{min}, p_{max}]$ of this parameter. We sample this range and select $p = p_{min} + 9.2\xi$, where $\xi = (p_{max} - p_{min})/10$. In the DSets algorithm, $s(x, y) = exp(-d(x, y)/\sigma)$ is used to evaluate the data similarity, and we manually select $\sigma = 10\overline{d}$ to obtain the best overall result, with $\overline{d}$ denoting the mean pairwise distance. With the DP-c and DP-G algorithms the parameter d_c is determined by including 1.1% and 2.0% of data into the neighborhood for DP-c and DP-G, respectively. We report the clustering results of these algorithms in Table 1.

We firstly look at the comparison between the original DP algorithms DP-c, DP-G and our algorithm. On D31 and R15 datasets, both DP-c and DP-G algorithms generate very good results, and our algorithm performs as well as these two. On the Compound, Spiral, Flame, Wine and Iris datasets, our algorithm compares favorably with the two algorithms. Only on the Aggregation dataset the two DP algorithms outperform ours evidently. These comparisons demonstrate the effectiveness of our improvements to the original DP algorithm.

Table 1. Clustering results (NMI) of some algorithms.

	k-means	NCuts	DBSCAN	AP	[19]	DSets	[11]	DP-c	DP-G	Ours
Aggregation	0.85	0.76	0.92	0.82	0.70	0.79	0.89	0.98	**0.99**	0.88
Compound	0.72	0.67	0.89	0.81	0.55	0.76	**0.92**	0.79	0.73	0.77
Spiral	0.00	0.00	0.71	0.00	0.00	0.32	0.66	0.36	**1.00**	**1.00**
D31	0.92	**0.96**	0.84	0.59	0.90	0.90	0.67	**0.96**	**0.96**	**0.96**
R15	0.96	**0.99**	0.87	0.74	0.94	0.86	0.91	0.98	**0.99**	**0.99**
Flame	0.39	0.44	0.83	0.57	0.30	0.50	0.90	**1.00**	0.41	**1.00**
Wine	0.43	0.36	0.38	0.39	**0.87**	0.77	0.43	0.61	0.71	0.74
Iris	0.74	0.76	0.75	0.79	0.73	0.64	0.56	0.66	0.66	**0.86**
mean	0.63	0.62	0.77	0.59	0.62	0.69	0.74	0.79	0.81	**0.90**

Comparatively, our algorithm is shown as the best-performing or near-best-performing one on 5 out of the 8 datasets, and our algorithm generates the best overall result. Especially on the Spiral dataset, on which k-means, NCuts and AP fail completely in clustering, our algorithm generate the perfect result. Even if our algorithm is outperformed by some algorithms on Aggregation, Compound and Wine datasets evidently, it is always among the 5 best-performing algorithms. These observations indicate that our algorithm has nice generality and performs well on various types of datasets.

5 Conclusions

An enhanced cluster center identification criterion and a new density kernel are presented to improve the DP clustering algorithm in this paper. By treating local density peaks as candidates of cluster centers, the DP algorithm uses local density and the distance to the nearest data of greater local density to represent the data and identify cluster centers. By studying the cluster center identification process, we find that local density is not very effective in strengthening the specificity of cluster centers. We introduce the concept of subordinates and present an alternative criterion to local density based on the subordinates. Furthermore, we make use of the average distance to neighboring data to evaluate the local density, in an endeavor to overcome the drawbacks of the cutoff and Gaussian kernels. Experiments show that the new criterion strengthens the specificity of cluster centers, and our algorithm performs well in comparison with some commonly used and recently proposed algorithms.

Acknowledgment. This work is supported in part by the National Natural Science Foundation of China under Grant No. 61473045, and by the Natural Science Foundation of Liaoning Province under Grant No. 20170540013 and 20170540005.

References

1. Achtert, E., Böhm, C., Kröger, P.: DeLi-Clu: boosting robustness, completeness, usability, and efficiency of hierarchical clustering by a closest pair ranking. In: Ng, W.-K., Kitsuregawa, M., Li, J., Chang, K. (eds.) PAKDD 2006. LNCS (LNAI), vol. 3918, pp. 119–128. Springer, Heidelberg (2006). https://doi.org/10.1007/11731139_16
2. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: Optics: ordering points to identify the clustering structure. In: ACM SIGMOD International Conference on Management of Data, pp. 49–60 (1999)
3. Brendan, J.F., Delbert, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007)
4. Chang, H., Yeung, D.Y.: Robust path-based spectral clustering. Pattern Recogn. **41**(1), 191–203 (2008)
5. Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. **17**(8), 790–799 (1995)
6. Daszykowski, M., Walczak, B., Massart, D.L.: Looking for natural patterns in data: part 1. density-based approach. Chemometr. Intell. Lab. Syst. **56**(2), 83–92 (2001)
7. Ester, M., Kriegel, H.P., Sander, J., Xu, X.W.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference on Knowledge Discovery and Data Mining, pp. 226–231 (1996)
8. Evanno, G., Regnaut, S., Goudet, J.: Detecting the number of clusters of individuals using the software structure: a simulation study. Mol. Ecol. **14**(8), 2611–2620 (2005)
9. Fu, L., Medico, E.: Flame, a novel fuzzy clustering method for the analysis of dna microarray data. BMC Bioinf. **8**(1), 1–17 (2007)
10. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. ACM Trans. Knowl. Discov. Data **1**(1), 1–30 (2007)
11. Hou, J., Gao, H., Li, X.: DSets-DBSCAN: a parameter-free clustering algorithm. IEEE Trans. Image Process. **25**(7), 3182–3193 (2016)
12. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn. **52**(1–2), 91–118 (2003)
13. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Advances in Neural Information Processing Systems, pp. 849–856 (2002)
14. Pavan, M., Pelillo, M.: Dominant sets and pairwise clustering. IEEE Trans. Pattern Anal. Mach. Intell. **29**(1), 167–172 (2007)
15. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344**, 1492–1496 (2014)
16. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 167–172 (2000)
17. Veenman, C.J., Reinders, M., Backer, E.: A maximum variance cluster algorithm. IEEE Trans. Pattern Anal. Mach. Intell. **24**(9), 1273–1280 (2002)
18. Zahn, C.T.: Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Trans. Comput. **20**(1), 68–86 (1971)
19. Zhu, X., Loy, C.C., Gong, S.: Constructing robust affinity graphs for spectral clustering. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1450–1457 (2014)