

Dominant Set Based Density Kernel and Clustering

Jian Hou^{1,2(✉)} and Shen Yin³

¹ College of Engineering, Bohai University, Jinzhou 121013, China
dr.houjian@gmail.com

² ECLT, Università Ca' Foscari Venezia, 30124 Venezia, Italy

³ Research Institute of Intelligent Control and Systems,
Harbin Institute of Technology, Harbin 150001, China

Abstract. The density peak based clustering algorithm has been shown to be a potential clustering approach. The key of this approach is to isolate and identify cluster centers by estimating the local density of data appropriately. However, existing density kernels are usually dependent on user-specified parameters evidently. In order to eliminate the parameter dependence, in this paper we study the definition of dominant set, which is a graph-theoretic concept of a cluster. As a result, we find that the weights of data in a dominant set provides a non-parametric measure of data density. Based on this observation, we then present an algorithm to estimate data density without parameter input. Experiments on various datasets and comparison with other density kernels demonstrate the effectiveness of our algorithm.

Keywords: Density peak · Clustering · Dominant set · Density kernel

1 Introduction

Data clustering has wide application in pattern recognition, image analysis and fault diagnosis [14, 15]. In the past decades, much efforts were devoted to data clustering and various clustering algorithms have been proposed. Unfortunately, in applying these algorithms to real clustering tasks, there are still many problems to be solved. The k-means-like algorithm, NCuts [12] and the general spectral clustering algorithms [17] uses as input the number of clusters and their results rely on the parameters heavily. In addition, these algorithms tend to generate clusters of spherical shapes, and the results are also influenced by cluster center initialization. The DBSCAN [5], AP [1] and DSets [10] algorithms are able to determine the number of cluster automatically. However, all these three algorithms have their own problems. The DBSCAN algorithm depends on two parameters *Eps* and *MinPts* for density estimation. Generally, the other density based clustering algorithms are also dependent on user-specified density

J. Hou—This work is supported in part by the National Natural Science Foundation of China under Grant No. 61473045 and by China Scholarship Council.

parameters. The AP algorithm must be fed the preference value of each data, which impacts on clustering results significantly. The DSets algorithm are built on the pairwise similarity matrix of the data to be clustered and is parameter independent in itself. However, in the case that data are represented as feature vectors, the estimation of data similarity usually introduces one or more similarity parameters, which have been found to impact on the clustering results [7, 8]. Besides, both the AP and DSets algorithms have the tendency to generate spherical clusters only. These observations show that it is still necessary to explore new clustering algorithms, although there are already a vast amount of clustering algorithms in the literature.

Our work in this paper is on the basis of the density peak (DP) based clustering algorithm proposed in [11]. The DP algorithm is based on the assumption that cluster centers are density peaks and they are relatively far from each other. With the local density ρ_i of each data i and the distance δ_i to the nearest neighbor with higher density to represent the data in a decision graph, it is found that the cluster centers are with both high ρ and high δ , whereas the non-center data are with either small ρ or small δ . As a result, the cluster centers are isolated from non-centered data and it is relatively easy to differentiate between two kinds of data. By assuming that the label of one data is the same as that of its nearest neighbor with higher density, all the non-center data can be grouped into clusters sequentially.

Local density calculation is the key of the DP algorithm as it determines ρ , δ and then the cluster centers. In [11] the authors use cut-off and Gaussian kernels, both of which involve a cut-off distance d_c . Although [11] presents an empirical method to calculate the range of d_c , we have found that the clustering results vary significantly with different d_c 's in this range. In order to solve this problem, in this paper we present a non-parametric density kernel based on the DSets algorithm. One important feature of the DSets algorithm is that each data in a cluster is assigned a weight. Our study of the dominant set definition shows that this weight reflects the relationship of the data with all the others, and can be viewed as a measure of the local density. By calculating the pairwise data similarity matrix properly, all the data can be included in one single cluster, and therefore the weights of all the data can be obtained with the DSets algorithm. We show that this process can be accomplished independent of user-specified parameters. The effectiveness of our algorithm is demonstrated in experiments and comparisons with other density kernels.

2 Density Peak Clustering

2.1 The DP Algorithm

The DP algorithm is proposed based on the following observations. First, cluster centers are usually the density peaks in the neighborhood. This means that compared with non-center data, cluster centers have relatively large local density ρ . Second, in practice few data are with the same local density, therefore the distance δ of one data to its nearest neighbor with higher density is usually

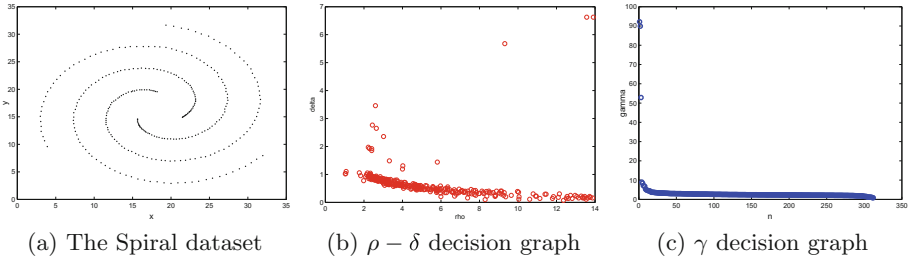


Fig. 1. The Spiral dataset and two decision graphs with the Gaussian kernel.

not large. In contrast, cluster centers are surrounded by data with lower density, therefore their δ 's are relative large. In summary, the cluster centers usually have both large ρ 's and large δ 's, whereas non-center data have either small ρ 's or small δ 's. This difference between cluster centers and non-center data makes it possible to isolate and identify cluster centers from non-center data. Then based on the assumption that the label of one data is the same as that of its nearest neighbor with higher density, the non-center data can be grouped into clusters. Although this assumption has no theoretic ground, it is consistent with human intuition and works well in experiments.

From the above description we see that the key of the DP algorithm is the calculation of local density. While the local density can be estimated in different ways, existing approaches usually involve user-specified parameters, which may influence the density values and then the clustering results. For example, the cutoff kernel and Gaussian kernel is used in [11] to calculate the local density, and both kernels involve the cutoff distance d_c . The cutoff kernel measures the density by the number of data in the neighborhood of radius d_c , and the Gaussian kernel uses d_c as the decay parameter. After the local density ρ 's are calculated, the distance δ_i is obtained by

$$\delta_i = \min_{j \in S, \rho_j > \rho_i} d_{ij}. \quad (1)$$

With the ρ 's and δ 's of all the data, we use a $\rho - \delta$ decision graph to illustrate the relationship of the cluster centers and non-center data in the $\rho - \delta$ space. Taking the Spiral dataset [3] for example, we show the $\rho - \delta$ decision graph in Fig. 1(b). For space reason, here we use only the Gaussian kernel and d_c is calculated by including 1.6% of all the data in the neighborhood.

It is evident in Fig. 1(b) there are three data with both large ρ 's and large δ 's, and they are presented as the outliers of the set of data. Obviously these three data are the centers of the three clusters. Considering that identifying cluster centers with the $\rho - \delta$ decision graph involves two thresholds, [11] then proposes to use $\gamma = \rho\delta$ as the single criterion of cluster center selection. We sort the data in the decreasing order according to their γ 's and obtain the γ decision graph in Fig. 1(c), where the three cluster centers with large γ 's can be recognized relatively easily.

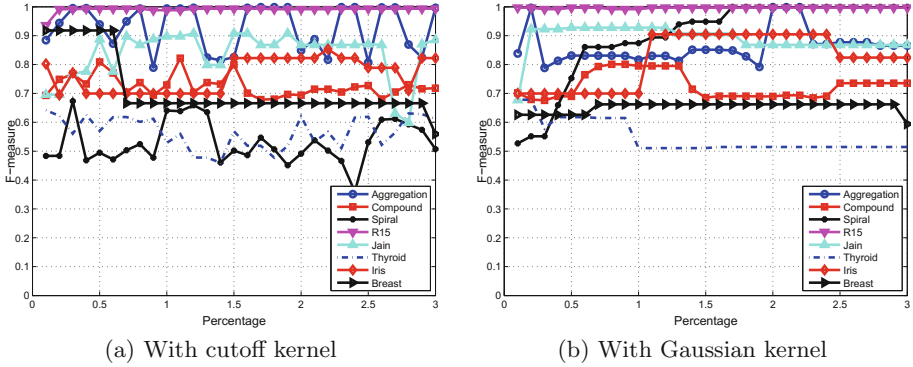


Fig. 2. The influence of the percentage in calculating d_c on the clustering results.

In general, while the $\rho-\delta$ decision graph and γ 's decision graph are helpful in identifying cluster centers, it is still quite difficult to find out the correct cluster centers automatically. In this paper we assume the number of clusters, N , is determined beforehand, and use the N largest γ 's to identify the cluster centers.

2.2 The Problems

In both the cutoff and Gaussian kernels the parameter d_c needs to be specified, and it is suggested in [11] to determine d_c such that 1% to 2% of all data are included in the neighborhood on average. However, we have found that with both kernels, different values of d_c in this range causes significant variance in the clustering results. In addition, the best results may not be obtained with d_c in this range. In fact, we show how the clustering results vary with the percentages used to calculate d_c in Fig. 2, where we use F-measure to evaluate the clustering results. Eight datasets, namely Aggregation [6], Compound [16], Spiral, R15 [13], Jain [9] and three UCI datasets Thyroid, Iris and Breast, are used in experiments.

Figure 2 indicates that the percentage and the parameter d_c has a significant influence on the clustering results. Unfortunately, we cannot arrive at any useful conclusion as to the appropriate range of d_c from Fig. 2. In addition, it is not clear how the clustering results are correlated with d_c . Consequently, it is very difficult to determine the appropriate d_c .

3 Our Algorithm

In order to solve the parameter dependence problem of existing density kernels, in this paper we present a non-parametric density by making use of the nice properties of the DSets algorithm. In this section we firstly introduce the dominant set definition, and then present in details how the DSets algorithm can be used to calculate the local density.

3.1 Dominant Set

In order to derive the non-parametric density kernel based on the dominant set, we firstly present the definition of dominant set briefly. The details of the definition can be found in [10].

We use S to denote the set of data for clustering, and $A = (a_{ij})$ to represent the pairwise similarity matrix. With D as a non-empty subset of S and $i \in D$, $j \notin D$, we measure the relationship of j and i by

$$\phi_D(i, j) = a_{ij} - \frac{1}{|D|} \sum_{k \in D} a_{ik}, \quad (2)$$

with $|D|$ denoting the number of data in D . Then we define

$$w_D(i) = \begin{cases} 1, & \text{if } |D| = 1, \\ \sum_{l \in D \setminus \{i\}} \phi_{D \setminus \{i\}}(l, i) w_{D \setminus \{i\}}(l), & \text{otherwise.} \end{cases} \quad (3)$$

With this key variable and $W(D) = \sum_{i \in D} w_D(i)$, the formal definition of dominant set can be presented as follows. A subset D such that $W(T) > 0$, for all non-empty $T \subseteq D$ is called a dominant set if

1. $w_D(i) > 0$, for all $i \in D$.
2. $w_{D \cup \{i\}}(i) < 0$, for all $i \notin D$.

In [10] the authors show that a dominant set can be extracted with game dynamics, e.g., replicator dynamics, developed in evolutionary game theory. Specifically, we use $x \in R^n$ to denote the weights of the data, which can be obtained by replicator dynamics. In this paper we adopt the more efficient dynamics proposed in [2]. It is shown that this weight vector corresponds to the weighted characteristic vector x^D of a dominant set D , which is defined as

$$x_i^D = \begin{cases} \frac{w_D(i)}{W(D)}, & \text{if } i \in D, \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

In other words, after we obtain the weight vector, the data with positive weights form a dominant set. In extracting a dominant set, the weights of all the data for clustering can be initialized to $\frac{1}{n}$. The dominant sets can be obtained sequentially in a peeling-off manner [10].

3.2 Non-parametric Density Kernel

From the last subsection we observe that in a dominant set, each data i is assigned a weight equaling to $\frac{w_D(i)}{W(D)}$. On the other hand, Eq. (3) indicates that $w_D(i)$ measures the similarity between i and the other data, and a large $w_D(i)$ means that i has a high overall similarity with other data. It is evident that if i is in the central area of a dominant set, then it is likely that $w_D(i)$ is large and

i has a large weight. In contrast, one data i in the border area of a dominant set tends to have a small weight. Since the weights of data in a dominant set can be used to differentiate between the data in central and border areas, they can be treated as the data density in the DP algorithm. In this sense, we can make use of the dominant sets algorithm to calculate the density, and therefore treat dominant set as a density kernel.

However, in applying this density kernel to the DP algorithm, there are still two problems to be solved. First, while the dominant set extraction uses as input only the pairwise similarity matrix and no parameters are involved, the calculation of data similarity usually introduces parameters. For example, the commonly used similarity measure $s(i, j) = \exp(-d(i, j)/\sigma)$ introduces the parameter σ . Second, in the case that there are more than one clusters in the dataset and the dynamics proposed in [2] are used, there will be some data with zero weights. These data with identical density will influence the clustering of non-center data negatively. By studying the definition of dominant set, in the following we show how to solve these two problems.

The definition of $w_D(i)$ in Eq. (3) indicates that a large $w_D(i)$ corresponds to large similarities between i and other data. Then the dominant set definition states that each data in a dominant set has a positive $w_D(i)$. This is equivalent to saying that each data in a dominant set are similar to all the others. As a result, the dominant set definition imposes a high requirement on the internal similarity in a dominant set. With a fixed dataset, the variance of σ results in the change of similarity value. A small σ leads to small similarity values, which further result in a large amount of small dominant sets. In contrast, a large σ corresponds to large similarity values and then a small number of large dominant sets. By adopting a sufficiently large σ , we can group all the data into a dominant set, and therefore assign non-zero weights to all the data. Although σ influences the similarity values, it does not change the magnitude ordering of these similarity values, and therefore has no influence on the ordering of data weights. Consequently, the value of σ does not impact on the DP clustering results, only if all data are assigned positive weights.

In practice, if σ is too large, many large similarity values may become identical due to limited digits after decimal. Therefore we use the following algorithm to determine the σ and generate the density used in the DP algorithm. With \bar{d} denoting the average of pairwise distances, we build a list composed of $\bar{d}, 10\bar{d}, 50\bar{d}, 100\bar{d}, 200\bar{d}, \dots$. Given a dataset, we assign σ with the values in the list from small ones to large ones, until all the resulted data weights are greater than zero.

4 Experiments

We test the proposed density kernel in experiments on the eight datasets, and compare the results with those from the cutoff kernel and Gaussian kernel. In addition, we also compare with some other algorithms, including k-means, DBSCAN, NCuts, AP and SPRG [17]. With k-means, NCuts and SPRG, we set

Table 1. Clustering results (F-measure) comparison on eight datasets.

	k-means	NCuts	DBSCAN	AP	SPRG	DP-cutoff	DP-Gaussian	Ours
Aggregation	0.83	0.99	0.80	0.82	0.73	0.99	0.99	0.94
Compound	0.68	0.70	0.88	0.77	0.64	0.82	0.69	0.81
Spiral	0.35	0.58	1.00	0.35	0.37	0.64	1.00	1.00
R15	0.82	0.99	0.77	0.54	0.93	0.99	0.99	0.95
Jain	0.79	0.63	0.87	0.57	0.86	0.90	0.87	1.00
Thyroid	0.83	0.64	0.68	0.52	0.97	0.55	0.51	0.72
Iris	0.89	0.93	0.77	0.93	0.87	0.70	0.90	0.78
Breast	0.96	0.64	0.87	0.82	0.97	0.67	0.66	0.69
Average	0.77	0.76	0.83	0.66	0.79	0.78	0.83	0.86

Table 2. Clustering results (Jaccard index) comparison on eight datasets.

	k-means	NCuts	DBSCAN	AP	SPRG	DP-cutoff	DP-Gaussian	Ours
Aggregation	0.64	0.98	0.67	0.71	0.49	0.98	0.99	0.87
Compound	0.46	0.46	0.84	0.69	0.42	0.71	0.47	0.74
Spiral	0.20	0.30	1.00	0.20	0.20	0.39	1.00	1.00
R15	0.65	0.99	0.42	0.25	0.83	0.96	0.99	0.86
Jain	0.53	0.42	0.91	0.29	0.63	0.71	0.65	1.00
Thyroid	0.64	0.40	0.57	0.29	0.90	0.29	0.29	0.58
Iris	0.69	0.79	0.59	0.77	0.66	0.51	0.73	0.60
Breast	0.87	0.39	0.78	0.56	0.89	0.48	0.40	0.54
Average	0.59	0.59	0.72	0.47	0.63	0.63	0.69	0.77

the required number of clusters as ground truth and report average results of 10 runs. With DBSCAN, the *MinPts* is set as 2, manually selected from 1 to 10, and *Eps* is determined based on *MinPts* [4]. For AP, the required preference value is manually selected to be $p_{min} + 9.2step$, where $step = (p_{max} - p_{min})/10$ and $[p_{min}, p_{max}]$ is the preference value range calculated with the method proposed by the authors of [1]. In DP algorithm with the cutoff and Gaussian kernel, the percentage of data used to calculate d_c is set as 1.1% and 2.0%, respectively, both of which are manually selected from 1, 1.1, 1.2, \dots , 2.0. The comparison of these algorithms are presented in Tables 1 and 2, where F-measure and Jaccard index are used to evaluate the clustering results. The comparison indicates that in terms of average clustering quality, our non-parameter kernel performs better than the cutoff and Gaussian kernels, and our algorithm also outperforms some other algorithms with carefully selected parameters.

5 Conclusions

In this paper we present a non-parametric density kernel to be used in the density peak based clustering algorithm. We study the dominant set definition and

propose to treat the extraction of dominant set as a density kernel which is independent of parameters. We compare with the cutoff and Gaussian kernels in the DP algorithm and also some other clustering algorithms to illustrate the effectiveness of the proposed density kernel. One problem with the proposed density kernel is the relatively high computation load involved in similarity calculation and density calculation, which will be studied in our future work.

References

1. Brendan, J.F., Delbert, D.: Clustering by passing messages between data points. *Science* **315**, 972–976 (2007)
2. Bulo, S.R., Pelillo, M., Bomze, I.M.: Graph-based quadratic optimization: a fast evolutionary approach. *Comput. Vis. Image Underst.* **115**(7), 984–995 (2011)
3. Chang, H., Yeung, D.Y.: Robust path-based spectral clustering. *Pattern Recognit.* **41**(1), 191–203 (2008)
4. Daszykowski, M., Walczak, B., Massart, D.L.: Looking for natural patterns in data: Part 1. Density-based approach. *Chemom. Intell. Lab. Syst.* **56**(2), 83–92 (2001)
5. 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)
6. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. *ACM Trans. Knowl. Discov. Data* **1**(1), 1–30 (2007)
7. Hou, J., Gao, H., Li, X.: DSets-DBSCAN: a parameter-free clustering algorithm. *IEEE Trans. Image Process.* **25**(7), 3182–3193 (2016)
8. Hou, J., Liu, W., Xu, E., Cui, H.: Towards parameter-independent data clustering. *Pattern Recognit.* **60**, 25–36 (2016)
9. Jain, A.K.: Data clustering: user's dilemma. In: Perner, P. (ed.) *MLDM 2007*. LNCS, vol. 4571, pp. 1–1. Springer, Heidelberg (2007). doi:[10.1007/978-3-540-73499-4_1](https://doi.org/10.1007/978-3-540-73499-4_1)
10. Pavan, M., Pelillo, M.: Dominant sets and pairwise clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(1), 167–172 (2007)
11. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014)
12. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8), 167–172 (2000)
13. Veenman, C.J., Reinders, M., Backer, E.: A maximum variance cluster algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(9), 1273–1280 (2002)
14. Yin, S., Gao, H., Qiu, J., Kaynak, O.: Descriptor reduced-order sliding mode observers design for switched systems with sensor and actuator faults. *Automatica* **76**, 282–292 (2017)
15. Yin, S., Gao, H., Qiu, J., Kaynak, O.: Fault detection for nonlinear process with deterministic disturbances: a just-in-time learning based data driven method. *IEEE Trans. Cybern.* (2016). doi:[10.1109/TCYB.2016.2574754](https://doi.org/10.1109/TCYB.2016.2574754)
16. Zahn, C.T.: Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.* **20**(1), 68–86 (1971)
17. 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)