Clustering by Searching Density Peaks via Local Standard Deviation

Juanying Xie(✉), Weiliang Jiang, and Lijuan Ding

School of Computer Science, Shaanxi Normal University,
Xi'an 710062, People's Republic of China
xiejuany@snnu.edu.cn

Abstract. To solve the problem of DPC (Clustering by fast search and find of Density Peaks) that it cannot find the cluster centers coming from sparse clusters, a new clustering algorithms is proposed in this paper. The proposed clustering algorithm uses the local standard deviation of point i to define its local density ρ_i, such that all the cluster centers no matter whether they come from dense clusters or sparse clusters will be found as the density peaks. We named the new clustering algorithm as SD_DPC. The power of SD_DPC was tested on several synthetic data sets. Three data sets comprise both dense and sparse clusters with various number of points. The other data set is a typical synthetic one which is often used to test the performance of a clustering algorithm. The performance of SD_DPC is compared with that of DPC, and that of our previous work KNN-DPC (K-nearest neighbors DPC) and FKNN-DPC (Fuzzy weighted K-nearest neighbors DPC). The experimental results demonstrate that the proposed SD_DPC is superior to DPC, KNN-DPC and FKNN-DPC in finding cluster centers and the clustering of a data set.

Keywords: Clustering · Density peaks · Local standard deviation · DPC · KNN-DPC · FKNN-DPC · SD_DPC

1 Introduction

Clustering analysis is to discover the group structure of a data set, and disclose the knowledge, patterns and rules hidden in the data [5,7,14,16]. It is an unsupervised learning process and implemented by grouping similar objects into same clusters and dissimilar ones in other clusters [1,3–5,10,12,13,16]. Its applications range from astronomy to bioinformatics, bibliometrics, biomedical, pattern recognition [1,3–5,11,15]. With the emerging of big data from various areas in the real world, there have been more and more experts focusing on studying the clustering techniques to try to understand and summarize the complex data automatically as far as possible and find the potential knowledge and rules and patterns embedded in the big data without any previous domain knowledge [1–3,9,11,15].

There are many kinds of clustering algorithms such as partitioning, hierarchical, density-based, etc. [1,3–5,12,13,16]. The novel density-based clustering

H. Yin et al. (Eds.): IDEAL 2017, LNCS 10585, pp. 295–305, 2017.
https://doi.org/10.1007/978-3-319-68935-7_33

algorithm was proposed by Alex Rodríguez and Alessandro Laio in [1]. The clustering algorithm can find the clustering of a data set by finding the density peaks as cluster centers and assign each point except for the density peaks to its nearest neighbor with higher density. We call the clustering algorithm as DPC for short. DPC is powerful except for its weaknesses in calculating the densities of points and its one step assignment strategy [1,12,13]. There are several advanced density peaks based clustering algorithms [1,8,12,13]. The K nearest neighbors based density peaks finding clustering (KNN-DPC) [12] and the fuzzy weighted K nearest neighbor based density peaks finding clustering (FKNN-DPC) [13] were proposed to remedy the deficiencies of DPC. The extensive experiments demonstrated the excellent performance of KNN-DPC and FKNN-DPC. However, because DPC uses the arbitrary *cutoff* distance d_c to define the local density of a point, which makes the cluster centers from sparse clusters may not be found for they cannot become density peaks. KNN-DPC and FKNN-DPC use the K nearest neighbors of a point to define its local density, and can solve the problem of DPC to some extent, but they did not thoroughly solve it. It is the common phenomena that there are both dense and sparse clusters in a date set simultaneously, which make it difficult for the aforementioned clustering algorithms to find the proper cluster centers and even the clustering as well.

To try to let the cluster centers can become density peaks no matter they come from dense or sparse clusters by absorbing the distributive information of points in a data set as far as possible, we propose the heuristic clustering algorithm by adopting the local standard deviation of a point to define its local density because it is well known that the local standard deviation of a point embodies the information of how dense the local area is around the point. As a result we introduce the new local standard deviation based density peaks finding clustering algorithm named SD_DPC. We tested its power on some special synthetic data sets which are composed of dense and sparse clusters simultaneously. We also test SD_DPC on a typical synthetic dataset from [6]. The experimental results demonstrate that SD_DPC is powerful in finding the cluster center by finding the local standard deviation based density peaks and the clustering of a data set, and its performance is superior to DPC, KNN-DPC and FKNN-DPC.

This paper is organized as follows: Sect. 2 introduces the new SD_DPC in detail. Section 3 tests the power of it by special synthetic data sets, and compares its performance with that of DPC, KNN-DPC and FKNN-DPC in terms of several typical criteria for testing a clustering algorithm, namely clustering accuracy (Acc), adjusted mutual information (AMI), and adjusted rand index (ARI). Section 4 is some conclusions.

2 The Main Idea of the Proposed Clustering Algorithm

DPC [1] has become the hotspot in machine learning for its strong power in detecting cluster centers and exclude outliers and recognize the clusters with any arbitrary shapes and dimensions. The basic assumption in DPC is that the ideal cluster centers are always surrounded by neighbors with lower local density,

and they are at a relatively large distance from any other points with higher local density. To find the ideal cluster centers, DPC introduces the *local density metric* of point i in (1) and the *distance* δ_i of point i in (2).

$$\rho_i = |\{d_{ij} | d_{ij} < d_c\}| \tag{1}$$

$$\delta_i = \begin{cases} max_j \ \{d_{ij}\}, & \rho_i = max_j\{\rho_j\} \\ min(\{d_{ij} | \rho_j > \rho_i\}), & otherwise \end{cases} \tag{2}$$

where d_{ij} is Euclidean distance between points i and j, and d_c the *cutoff* distance given manually. We can see from (1) that the local density ρ_i of point i is the number of points j that are closer to i than d_c. DPC is robust to d_c for large data sets [1]. The definition in (2) disclose that δ_i is the maximum distance from point i to any other point j when point i has got the highest density, otherwise δ_i is the minimum distance between point i and any other point j with higher density [1]. It can be seen from (2) that δ_i is much larger than the nearest neighbor distance for the points with local or global maxima density. The points with anomalously large value of δ_i are to be chosen as cluster centers by DPC.

The most important contribution of DPC is that it proposed the idea of decision graph which is the collection of points (ρ_i,δ_i) in a 2-dimension space with ρ and δ to be x-axis and y-axis respectively. Cluster centers are the points with high δ and relatively high ρ, that is the cluster centers are the ones at the top right corner of the decision graph. The second innovation of DPC is its one step assignment that it assign each remaining points except for those density peaks to the same cluster as its nearest neighbor with higher density. This one step assignment contribution leads the DPC's efficient execution.

However, everything has two sides. The local density definition of DPC in (1) may results in the lower density for those cluster centers from sparse clusters for the arbitrary *cutoff* distance d_c. The very efficient one step assignment of DPC for remaining points may lead to the similar "Domino Effect" [12,13], that is once a point is assigned erroneously, then there may be many more points subsequently assigned incorrectly [12,13], especially in the case where there are several overlapping clusters.

$$\rho_i = \sum_{j \in KNN_i} exp(-d_{ij}) \tag{3}$$

KNN-PDC [12] and FKNN-DPC [13] introduced new definition of density ρ_i for point i by (3). The new definition can be used to any size of data set to calculate the density ρ_i of point i. The KNN_i in (3) means the data set composed of the points of the K-nearest neighbors of point i. Furthermore, KNN-DPC and FKNN-DPC respectively proposed their own two-step assignment strategy, where FKNN-DPC introduced fuzzy weighted K nearest neighbors theory in its second step of assignment strategy other than that of KNN-DPC only the K nearest neighbors theory is used in its two-step assignment strategy. It was demonstrated that both KNN-DPC and FKNN-DPC are superior to DPC [12, 13]. Due to the contribution of fuzzy weighted K-nearest neighbors that FKNN-DPC is more robust than KNN-DPC [13].

Although KNN-DPC and FKNN-DPC have been demonstrated very powerful in finding the patterns of data sets, they did not solve the problem of DPC thoroughly, such as finding the cluster centers from the sparse clusters by finding density peaks.

To detect the cluster centers from sparse clusters by finding the density peaks, we try to absorb the distribution information of points in a dataset as far as possible, so we introduce the new local density ρ_i for point i in (4) by introducing the local standard deviation of point i with the knowledge that the local standard deviation of a point embodying the local information of how dense the local area is around point i. The KNN_i in (4) is the same as that in (3), that is the set of K nearest neighbors of point i, and d_{ij} the Euclidean distance between data points i and j. Then the new local standard deviation based density peaks finding clustering algorithm is proposed in this paper and named as SD_DPC. The assignment strategy in SD_DPC is the same as that of DPC only to demonstrate that the local density definition ρ_i of point i in (4) can solve the problem existing in DPC that the cluster center of sparse cluster may not be detected by finding density peaks.

$$\rho_i = \frac{1}{\sqrt{\frac{1}{K-1}\sum_{j \in KNN_i} d_{ij}^2}} \tag{4}$$

Here are the main steps of SD_DPC.

step1 calculate the density ρ_i of point i in (4), and its distance δ_i in (2);

step2 plot all of points in the 2-dimensional space with their densities ρ and δ as their x and y coordinates respectively;

Table 1. Description of synthetic data sets

Datasets	Number of records	Number of attributes	Number of clusters
a3	7,500	2	50
dataset1	678	2	3
dataset2	10,900	2	8
dataset3	20,000	2	4

Table 2. The parameters for generating dataset1

Parameter	cluster1	cluster2	cluster3
Mean	[3, 2]	[8, 2]	[6, 5.5]
Covariance	$\begin{bmatrix} 0.7, & 0 \\ 0, & 0.7 \end{bmatrix}$	$\begin{bmatrix} 0.7, & 0 \\ 0, & 0.7 \end{bmatrix}$	$\begin{bmatrix} 1, 0 \\ 0, 1 \end{bmatrix}$
Number of points	300	300	78

Table 3. The parameters for generating dataset2

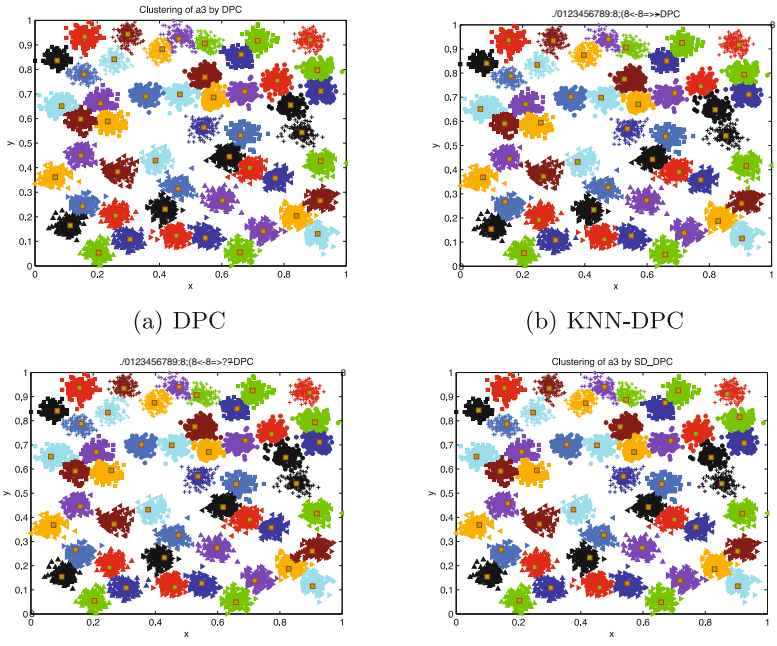
Parameter	cluster1	cluster2	cluster3	cluster4	cluster5	cluster6	cluster7	cluster8
Mean	[1, 8]	[1, 3]	[3, 6]	[8, 10]	[9, 2]	[13, 6]	[16, 1]	[16, 12]
Covariance	$\begin{bmatrix} 0.2, & 0 \\ 0, & 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.2, & 0 \\ 0, & 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.2, & 0 \\ 0, & 0.2 \end{bmatrix}$	$\begin{bmatrix} 1, & 0 \\ 0, & 1 \end{bmatrix}$	$\begin{bmatrix} 1, & 0 \\ 0, & 1 \end{bmatrix}$	$\begin{bmatrix} 1, & 0 \\ 0, & 1 \end{bmatrix}$	$\begin{bmatrix} 0.2, & 0 \\ 0, & 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.2, & 0 \\ 0, & 0.2 \end{bmatrix}$
Number of points	2,000	2,000	2,000	300	300	300	2,000	2,000

Table 4. The parameters for generating dataset3

Parameter	cluster1	cluster2	cluster3	cluster4
Mean	[2, 2]	[9, 2]	[6, 5.5]	$x \in [0.5,\ 12.5]$
Covariance	$\begin{bmatrix} 0.2, & 0 \\ 0, & 0.2 \end{bmatrix}$	$\begin{bmatrix} 1, & 0 \\ 0, & 1 \end{bmatrix}$	$\begin{bmatrix} 1.5, & 0 \\ 0, & 1.5 \end{bmatrix}$	$y \in [8,\ 9.5]$
Number of points	8,000	3,000	3,999	5,001

step3 select those points at the top right corner in the 2-dimensional space, that is density peaks with relatively higher densities and distances, as cluster centers of the data set;

step4 assign the remaining points except for the cluster centers to its nearest neighbor with higher density.



(a) DPC (b) KNN-DPC

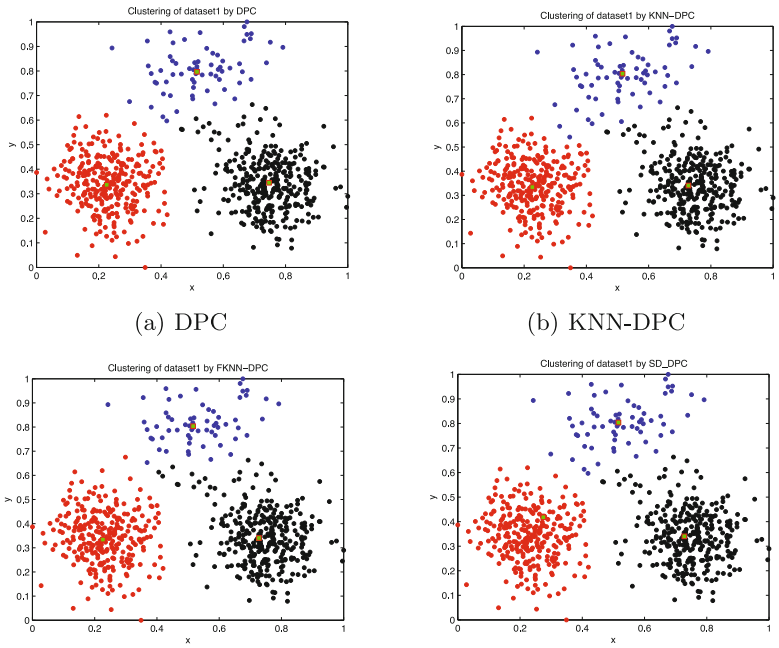
(c) FKNN-DPC (d) SD_DPC

Fig. 1. The clusterings of a3 by 4 clustering algorithms, respectively.

3 Experiments and Analysis

This section will display the experimental results of SD_DPC and the analysis. In order to test the power of SD_DPC, especially its power to detect the cluster centers from sparse clusters and find the clustering of a data set which has got both dense and sparse clusters simultaneously, we synthetically generated data sets with both dense and sparse clusters. In addition, we test the power of SD_DPC by the typical data set a3 from other reference. It should be explained that for the page limitation we cannot display more experimental results on any other typical bench mark data sets. The Subsect. 3.1 will display the data sets used in this paper. The experimental results of SD_DPC and the analysis are shown in Subsect. 3.2. We compared the performance of SD_DPC with that of DPC, KNN-DPC and FKNN-DPC. For the page limitation, the decision graph and some other experimental results such as the comparison with other clustering algorithms cannot be included in this paper.

We normalize the data using a min-max normalization given by (5), where x_{ij} is the value of attribute j of point i, and $min(x_j)$ and $max(x_j)$ are the minimum and maximum values of attribute j, respectively. The min-max normalization in (5) can preserve the original relationship in data [4], and reduce the influence on experimental results from different metrics for attributes and



(a) DPC (b) KNN-DPC

(c) FKNN-DPC (d) SD_DPC

Fig. 2. The clusterings of dataset1 by 4 clustering algorithms, respectively

reduce the runtime of algorithms' as well.

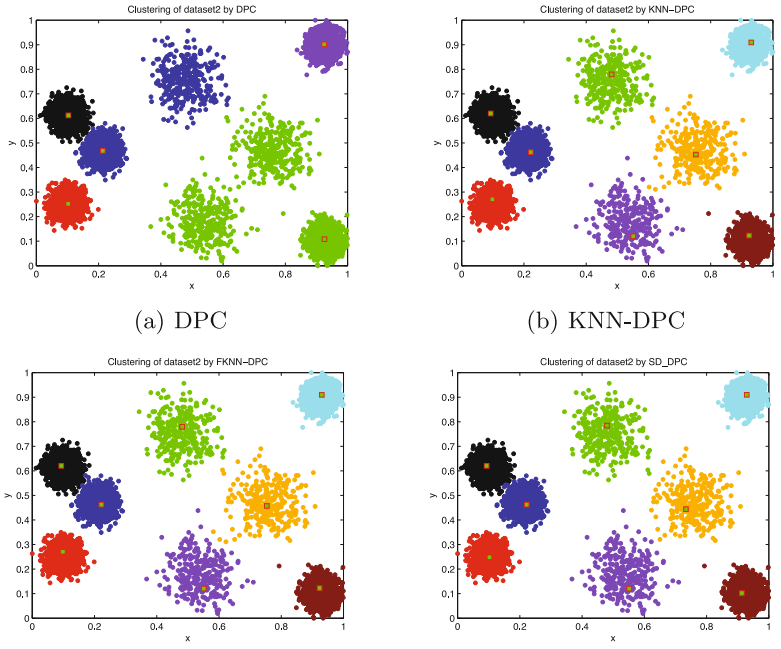
$$x_{ij} \leftarrow \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)} \tag{5}$$

3.1 Data Sets Description

Table 1 shows the informations of the synthetic data sets used in our experiments. The data set a3 comes from [6], and the other three data sets are synthetically generated. The parameters for generating the synthetical data sets are shown in Tables 2, 3 and 4, respectively. We designed these various size of data sets with dense and sparse clusters simultaneously only to test the ability and the scalability of SD_DPC in finding the cluster center from a sparse cluster and the clustering of a data set as well.

3.2 Results and Analysis

Figures 1, 2, 3 and 4 respectively display the clusterings of data sets from Table 1 by 4 clustering algorithms. The square point in each cluster is the cluster center detected by the related algorithm. Table 5 displays the quantity results of the



(a) DPC (b) KNN-DPC

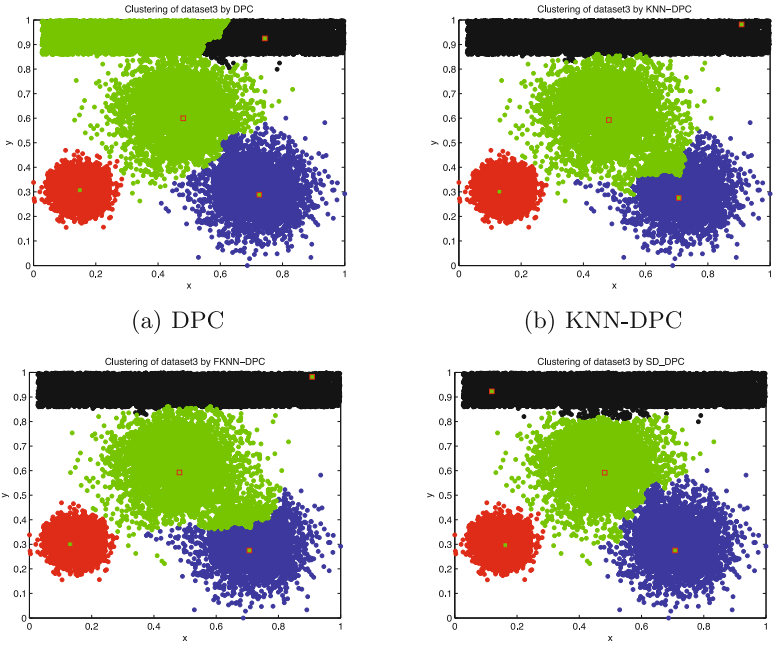
(c) FKNN-DPC (d) SD_DPC

Fig. 3. The clusterings of dataset2 by 4 clustering algorithms, respectively

aforementioned clustering algorithms in terms of Acc, AMI and ARI, and the number of clusters discovered by each algorithm and the number of clusters covering the cluster centers by F/P, where F refers to the number of cluster centers found by algorithms and P the number of clusters in which the cluster centers lie. The parameters pre-specified for each algorithm are also shown in Table 5 by *Par*.

From the clusterings of a3 and dataset1 by 4 clustering algorithms respectively shown in Figs. 1 and 2, we can say that all of the 4 clustering algorithms can detect all of the cluster centers and the clustering of a3 and dataset1. The difference between the clusterings by 4 algorithms only comes from the border points between clusters. The quantity comparison between the clustering results of 4 clustering algorithms on a3 and dataset1 will be displayed in Table 5 in terms of Acc, AMI and ARI.

The clusterings shown in Fig. 3 disclose that the cluster centers of three sparse clusters in dataset2 cannot be detected by DPC, while the other three algorithms can detect all the cluster centers by finding density peaks. However, the cluster centers found by 4 clustering algorithms are not completely same, and border points between cluster5 and cluster6 are grouped into different clusters by KNN-DPC, FKNN-DPC and SD_DPC. The clustering by SD_DPC is the



(a) DPC (b) KNN-DPC

(c) FKNN-DPC (d) SD_DPC

Fig. 4. The clusterings of dataset3 by 4 clustering algorithms, respectively

Table 5. Comparison of 4 clustering algorithms on 4 synthetic data sets.

Algorithm	a3					dataset1				
	Acc	AMI	ARI	F/P	Par	Acc	AMI	ARI	F/P	Par
DPC	**0.989**	**0.986**	0.977	50/50	1.25	0.978	0.896	0.946	3/3	5
KNN-DPC	**0.989**	**0.986**	0.978	50/50	10	**0.981**	**0.907**	**0.953**	3/3	12
FKNN-DPC	0.987	0.985	0.975	50/50	10	0.973	0.877	0.934	3/3	12
SD_DPC	**0.989**	**0.986**	**0.979**	50/50	6	0.978	0.896	0.946	3/3	7
Algorithm	dataset2					dataset3				
	Acc	AMI	ARI	F/P	Par	Acc	AMI	ARI	F/P	Par
DPC	0.917	0.864	0.895	5/5	2	0.848	0.810	0.779	4/4	2
KNN-DPC	**0.999**	**0.997**	**0.999**	8/8	10	0.959	0.908	0.931	4/4	12
FKNN-DPC	**0.999**	0.996	0.998	8/8	12	0.968	0.918	0.947	4/4	12
SD_DPC	**0.999**	**0.997**	**0.999**	8/8	13	**0.982**	**0.940**	**0.965**	4/4	8

best one compared to original pattern of dataset2, which is not displayed for the page limitations.

The clusterings of dataset3 by 4 clustering algorithms shown in Fig. 4 demonstrate the power of our SD_DPC. From the results in Fig. 4, it can be seen that DPC cannot detect the clustering of dataset3, while the other three algorithms can. KNN-DPC and FKNN-DPC are much better than DPC in finding the clustering of dataset3, but they are not as good as our SD_DPC for the mistakes they made in finding the cluster2 and cluster3 of dataset3. The detail quantity evaluation of the 4 clustering algorithms on dataset3 will be shown in Table 5.

The clustering results in Table 5 reveal that our SD_DPC has got the best performance among the 4 clustering algorithms while it was defeated by KNN-DPC on dataset1. Our previous study KNN-DPC is also a good clustering algorithm. DPC only has got comparable performance on a3.

The overall analysis demonstrate that our proposed SD_DPC is powerful in finding the cluster centers and the clustering of a data set no matter the clusters are dense or sparse and with any arbitrary shapes. So we can conclude that the proposed local density of a point based on its local standard deviation is valid.

4 Conclusions

This paper proposed to adopt the local standard deviation of point i to define its local density ρ_i, so that the distribution information of points in a data set can be absorbed as much as possible to overcome the problem of DPC which may not detect the cluster center of a sparse cluster by finding density peaks. As a consequence the heuristic clustering algorithms named SD_DPC has been introduced. The performance of SD_DPC was tested on several synthetic data sets and compared with that of DPC, KNN-DPC, FKNN-DPC. The experimental

results demonstrate that the proposed SD_DPC is superior to DPC, KNN-DPC and FKNN-DPC.

Acknowledgments. We are much obliged to those who provide the public data sets for us to use. This work is supported in part by the National Natural Science Foundation of China under Grant No. 61673251, is also supported by the Key Science and Technology Program of Shaanxi Province of China under Grant No. 2013K12-03-24, and is at the same time supported by the Fundamental Research Funds for the Central Universities under Grant No. GK201701006, and by the Innovation Funds of Graduate Programs at Shaanxi Normal University under Grant No. 2015CXS028 and 2016CSY009.

References

1. Alex, R., Alessandro, L.: Clustering by fast search and find of density peaks. Science **344**(6191), 1492–1496 (2014)
2. Dan, F., Melanie, S., Christian, S.: Turning big data into tiny data: constant-size coresets for k-means, PCA and projective clustering. In: Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, pp. 1434–1453. SIAM (2013), http://dl.acm.org/citation.cfm?id=2627817.2627920
3. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814), 972–976 (2007)
4. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann (2011)
5. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern Recogn. Lett. **31**(8), 651–666 (2010)
6. Karkkainen, I., Franti, P.: Dynamic local search for clustering with unknown number of clusters. In: Proceedings of the 16th International Conference on Pattern Recognition, vol. 2, pp. 240–243. IEEE (2002)
7. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Statistics, Oakland, CA, USA, pp. 281–297 (1967)
8. Mehmood, R., EI-AShram, S., Bie, R., Dawood, H., Kos, A.: Clustering by fast search and merge local density peaks for gene expression microarray data. Sci. Rep. **7**, 45602 (2017)
9. Tong, H., Kang, U.: Big data clustering. In: Aggarwal, C.C., Reddy, C.K. (eds.) Data Clustering: Algorithms and Applications, chap. 11, pp. 259–276. CRC Press (2013)
10. Von Luxburg, U., Williamson, R.C., Guyon, I.: Clustering: science or art? J. Mach. Learn. Res. Proc. Track **27**, 65–80 (2012)
11. Xie, J., Gao, H.: Statistical correlation and k-means based distinguishable gene subset selection algorithms. J. Softw. **25**(9), 2050–2075 (2014)
12. Xie, J., Gao, H., Xie, W.: K-nearest neighbors optimized clustering algorithm by fast search and nding the density peaks of a dataset. SCIENTIA SINICA Informationis **46**(2), 258–280 (2016)
13. Xie, J., Gao, H., Xie, W., Liu, X., Grant, P.W.: Robust clustering by detecting density peaks and assigning points based on fuzzy weighted k-nearest neighbors. Inf. Sci. **354**, 19–40 (2016)

14. Xie, J., Jiang, S., Xie, W., Gao, X.: An efficient global K-means clustering algorithm. J. Comput. **6**(2), 271–279 (2011)
15. Xie, J., Li, Y., Zhou, Y., Wang, M.: Differential feature recognition of breast cancer patients based on minimum spanning tree clustering and F-statistics. In: Yin, X., Geller, J., Li, Y., Zhou, R., Wang, H., Zhang, Y. (eds.) HIS 2016. LNCS, vol. 10038, pp. 194–204. Springer, Cham (2016). doi:10.1007/978-3-319-48335-1_21
16. Xu, R., Wunsch, D.I.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)