Neurocomputing 208 (2016) 210–217

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Clustering by fast search and find of density peaks via heat diffusion

Rashid Mehmood [a,b], Guangzhi Zhang [a], Rongfang Bie [a,*], Hassan Dawood [d], Haseeb Ahmad [c]

[a] College of Information Science and Technology, Beijing Normal University, Beijing 100875, China
[b] Department of Computer Science and Information Technology, University of Management Sciences and Information Technology, Kotli, AJK, Pakistan
[c] Information Security Center, State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, People's Republic of China
[d] Department of Computer Engineering, University of Engineering and Technology, Taxila, Pakistan

ARTICLE INFO

Article history:
Received 27 October 2015
Received in revised form
29 December 2015
Accepted 24 January 2016
Available online 7 June 2016

Keywords:
Clustering
Probability density estimation
Kernel density estimation
Heat equation

ABSTRACT

Clustering by fast search and find of density peaks (CFSFDP) is a novel algorithm that efficiently discovers the centers of clusters by finding the density peaks. The accuracy of CFSFDP depends on the accurate estimation of densities for a given dataset and also on the selection of the cutoff distance (d_c). Mainly, d_c is used to calculate the density of each data point and to identify the border points in the clusters. CFSFDP necessitates using different methods for estimating the densities of different datasets and the estimation of d_c largely depends on subjective experience. To overcome the limitations of CFSFDP, this paper presents a method for CFSFDP via heat diffusion (CFSFDP-HD). CFSFDP-HD proposes a nonparametric method for estimating the probability distribution of a given dataset. Based on heat diffusion in an infinite domain, this method accounts for both selection of the cutoff distance and boundary correction of the kernel density estimation. Experimental results on standard clustering benchmark datasets validate the robustness and effectiveness of the proposed approach over CFSFDP, AP, mean shift, and K-means methods.

1. Introduction

Clustering plays an important role in the fields of knowledge discovery and data mining. Clustering algorithms attempt to organize data into different disjoint categories, with more similar data points organized into the same cluster, while dissimilar data points are grouped into different clusters.

Clustering has been successfully applied in different fields such as bioinformatics [1–3], cyber security [4,5], image processing [6–11], astronomy [12], social networks [13,14] etc. Clustering algorithms are categorized into density based [15–20], hierarchical [21–24], partitioning [25,26], model based [27,28], and grid based [29] methods.

Density based clustering algorithms create arbitrary shapes of clusters even in the presence of noise in large spatial databases. They require minimum domain knowledge to cluster the datasets [16].

DBSCAN [18] is a popular density based clustering algorithm that discovers arbitrary shaped clusters. It is robust against noise, requires minimal input parameters, and scales well for large datasets. However, it is not fully deterministic for border points—cluster shapes depend upon input parameters—and it can become stuck on overlapping densities. A number of variants have been proposed to overcome these deficiencies, such as DBCLASD [16], OPTICS [17], ST-DBSCAN [19], and VDBSCAN [20].

Recently an algorithm implementing clustering by fast search and find of density peaks (CFSFDP) was proposed by Alex and Laio [30]. CFSFDP is based on two assumptions: the central point of a cluster has higher density compared to its neighbors, and the cluster center is relatively far from other cluster centers compared to its local data points. For each data point, i, CFSFDP computes a local density, ρ_i, and a distance, δ_i, relative to the nearest high density point. The effectiveness of CFSFDP depends greatly on the accurate estimation of density and the cutoff distance, d_c. d_c is an essential parameter for estimating the accurate density of a data point and identifying border points of a cluster.

The selection of d_c is based on a heuristic approach that the average number of neighbors in a dataset should only be 1–2% of the entire dataset. A better choice of selecting d_c is related to the user's observation with respect to the nature of the dataset. Therefore, CFSFDP faces some limitations in that it is hard for users to estimate the sensitive parameter, d_c; robust methods for calculating accurate densities are not available [31,32]; and different

* Corresponding author.
E-mail addresses: gulkhan007@gmail.com (R. Mehmood), zgz@mail.bnu.edu.cn (G. Zhang), rfbie@bnu.edu.cn (R. Bie), hasandawod@yahoo.com (H. Dawood), haseeb_ad@hotmail.com (H. Ahmad).

http://dx.doi.org/10.1016/j.neucom.2016.01.102

methods are required to estimate density based on the nature of the dataset.

To overcome the aforementioned issues, this paper proposes a new algorithm: CFSFDP-HD, where (1) heat-diffusion method [31] is introduced to estimate underlying density, (2) the sensitivity parameter, d_c, is simplified, and (3) the time parameter of heat diffusion is proposed to detect the border points of the clusters in an efficient way.

Background knowledge is presented in Section 2. Section 3 describes the kernel density estimation and our proposed method in detail. Experimental results are presented and discussed in Section 4. The conclusions and future work are stated in Section 5.

2. Background knowledge

CFSFDP has the ability to create arbitrary shaped clusters by fast search of cluster centers. It assumes that the cluster center is a highly dense data region and is positioned at a relatively large distance from other cluster centers compared to its local data points. For each data point, i, CFSFDP computes its local density, ρ_i, and distance, δ_i, to its nearest high density data point. The local density is

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

where

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

d_{ij} is the distance from data point i to data point j and d_c is the cutoff distance. ρ_i is equal to the number of data points that are closer than d_c to i. Thus, d_c is an essential parameter used to calculate the density of each data point. The effectiveness of CFSFDP depends greatly upon the appropriate choice of d_c. For small datasets, ρ_i can be affected by large statistical errors [30], in which the approach of [34,35] for estimating the density is recommended.

The distance, δ_i, of a data point to the nearest highly dense data point, $\max_{\rho_i}$, is calculated for the purpose of assigning i to the nearest cluster center,

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

Data points with high local or global density have the maximum value of δ. Hence, cluster centers are those points with high ρ and large δ compared to other points in the dataset. After computing ρ_i and δ_i for each data point, these statistics are plotted on a decision graph, as shown in Fig. 1.

In Fig. 1(a), 28 data points are shown with decreasing density order, and Fig. 1(b) shows the corresponding decision graph. Points 1 and 10 show high density with high δ, which is characteristic of cluster centers. Since points 26, 27, and 28 are isolated, they have high δ and low ρ, and can be considered as noise or outliers. Thus, using the decision graph, the expected cluster centers can be easily identified. After successful identification of the cluster centers, CFSFDP assigns the remaining data points to the nearest cluster center based on their δ values in a single round.

A border region is identified for each cluster and contains data points that are part of the underlying cluster and also fall within the d_c of another cluster. For these border points, CFSFDP finds the maximum density, ρ_b, within the border region of the underlying cluster, and those points with higher density than ρ_b are considered as cluster points, while the other data points are identified as cluster halo points and can be considered as noise.

3. Proposed technique

3.1. Kernel density estimation

Nonparametric density estimation is an important tool for statistical analysis of data. It is used to evaluate skewness, multimodality, summarizing Bayesian posteriors, discriminant analysis, and classification [31,36]. Nonparametric approaches are more flexible for modelling of datasets and are not affected by specification bias [36], in contrast to the classical approach [31]. Kernel density estimation (KDE) is the most commonly used nonparametric density estimation method [31]. The state-of-the-art method for estimating the density is to introduce a narrow Gaussian kernel (or alternatives), $\hat{f}_h(d_i)$, at each data point d_i and compute the integral of all kernel values over the entire dataset [37,38]. The KDE for identical and independent data points $\{d_1, d_2, d_3 \ldots, d_n\}$ drawn with an unknown probability density function (PDF) is

$$\hat{f}_h(d; h) = \frac{1}{n} \sum_{j=1}^{n} K_h(d - d_j) \tag{3}$$

The Gaussian kernel, $K\left(d, d_j; h\right)$ is normally used to estimate the density,

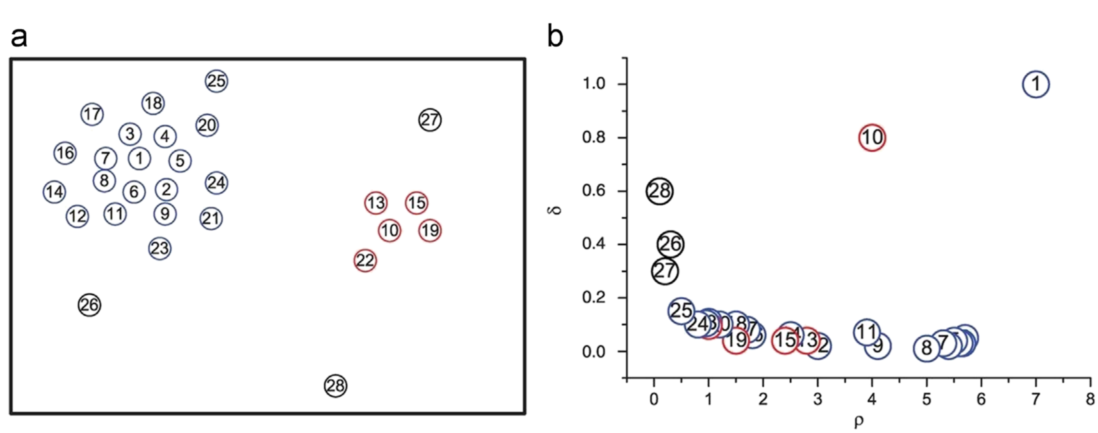

Fig. 1. CFSFDP in two dimensions. (a) Data points distribution. (b) Decision graph for data in (a). Different colors represent different clusters [30]. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 1
The detail description of datasets.

Dataset	Objects (n)	Dimensions (d)	Classes (k)	Sources
Point distributions	2000	2	5	[30]
Aggregation	788	2	7	[41]
flame	240	2	2	[42]
Path-based spiral	312	2	2	[43]
R15	600	2	15	[44]
D31	3100	2	31	[44]
Dim2	1650	2	9	[45]
Toys problem	300	2	3	[46]
A1	3000	2	20	[47]
Diamond	3000	2	9	[48]
S1	5000	2	15	[49]

$$K(d, d_j;\, h) = \frac{1}{\sqrt{2\pi h}} e^{-\frac{(d-d_j)^2}{2h}}, \tag{4}$$

where K is the kernel function scaled by $1/h$, and h is the bandwidth of the kernel function. The performance of Eq. (3) depends greatly upon the appropriate choice of h [39,40]. The mean integrated squared error (MISE) [31] is a well-studied criteria used to determine an optimal value of h,

$$MISE\{\hat{f}\}(h) = \mathbb{E}_f \int [\hat{f}(d;\, h) - f(d)]^2 dx. \tag{5}$$

The Gaussian KDE has some limitations, e.g., the sensitive parameter h (bandwidth) is difficult to select, boundary bias, and under or over smoothing.

3.2. Proposed method for density estimation

Rather than Eq. (2) or (3) for estimating the density of a given dataset, we propose to use the KDE via the heat diffusion method [31]. Heat diffusion method views the kernel density estimate as a unique solution to the diffusion partial differential equation, which evolves for a time t proportional to the kernel bandwidth h [31–33]. The interpretation of KDE via heat diffusion derives from the concept of the Weiner process, W, a continuous time stochastic process where the next stage is directly calculated by the previous state, such that

1. The preparatory probability is equally distributed through the d-dimensional data points $\{d_1, d_2, d_3, \ldots, d_n\}$.
2. The Gaussian kernel is used to estimate the transition probability from point d_i to d_j,

$$P_{\text{trnasition}(d,d_j;t)} = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(d-d_j)^2}{2t}} \tag{6}$$

The KDE is interpreted as the probability distribution function for this process at time t, which is similar to Eq. (3) with bandwidth h,

$$\hat{f}_t(d;\, t) = \frac{1}{2} \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(d-d_j)^2}{2t}}. \tag{7}$$

Eq. (7) is an iterative process, therefore the transition satisfies the diffusion partial differential equation (PDE),

$$\frac{\partial}{\partial t} \hat{f}(d;\, t) = \frac{1}{2} \frac{\partial^2}{\partial d^2} \hat{f}(d;\, t), \quad d \in D, \quad t > 0, \tag{8}$$

where $D \equiv \mathbb{R}$ and the initial condition $\hat{f}(d;\, 0) = \Delta(d)$, where $\Delta(d) = \frac{1}{n}\sum_{j=1}^{n} \delta(d - d_j)$ is an empirical density of dataset D, and $\delta(d - d_j)$ is the Dirac delta function that assigns point masses to all points of the dataset. When the domain has finite endpoints, Eq. (3) needs boundary correction. Therefore, within the PDE framework, we have to solve Eq. (8) over the finite domain with the initial condition $\hat{f}(d;\, 0) = \Delta(d)$ and the Neumann boundary condition,

$$\frac{\partial}{\partial t} \hat{f}(d;\, t)\Big|_{d=X_l} = \frac{\partial}{\partial t} \hat{f}(d;\, t)\Big|_{d=Xu} = 0, \tag{9}$$

where X_l and Xu are the lower and upper bounds of the domain, respectively. Considering the Neumann boundary condition and probability density with domain $[0, 1]$, the analytical solution of this PDE can be written in form of the theta kernel, θ, instead of

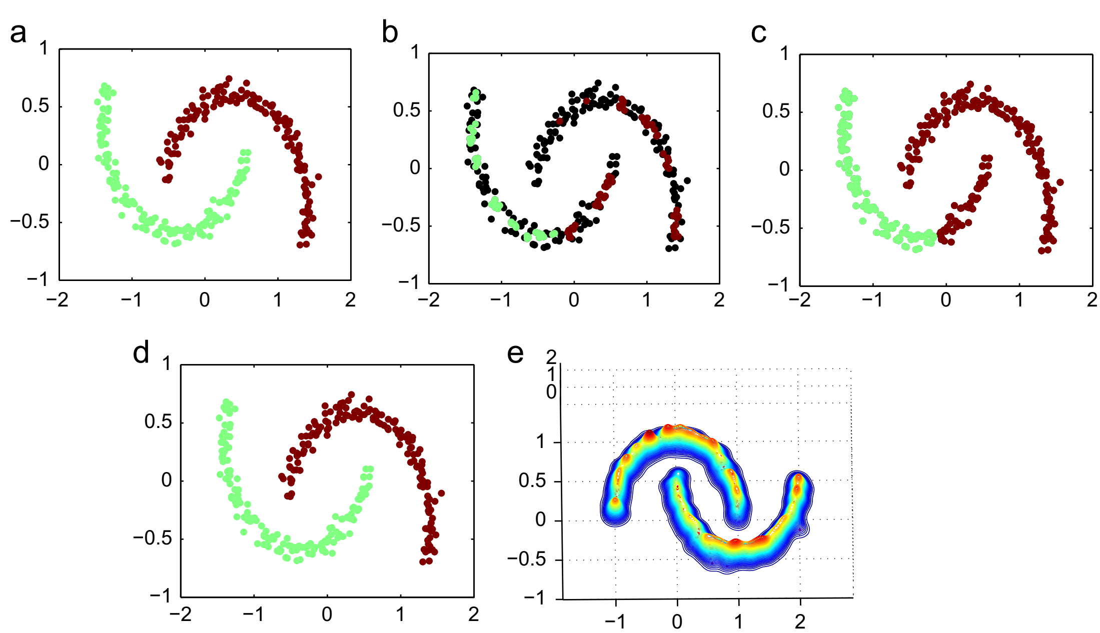

Fig. 2. CFSFDP created clusters of toys problem at different values of d_c and comparison with proposed method. (a) Presents toys problem synthetic dataset. (b) and (c) Show the CFSFDP clusters formed by considering d_c as 2% and 1% of the entire dataset, respectively. (d) Presents clusters analyzed using the CFSFDP-HD. (e) Shows estimated densities by proposed method in 3D space.

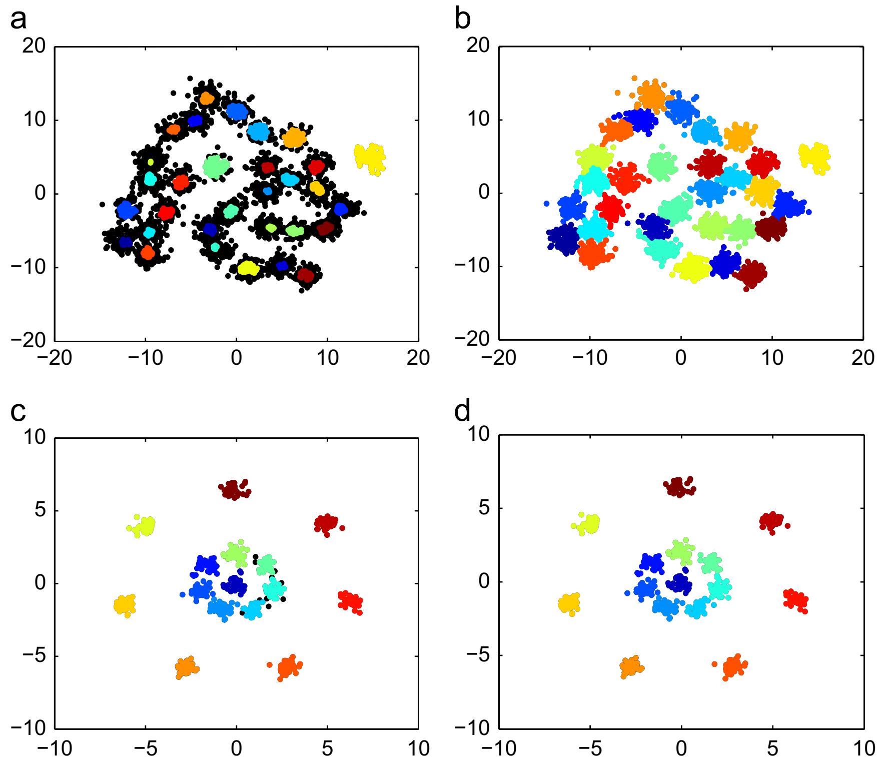

Fig. 3. Presents a comparison of clusters created by CFSFDP and CFSFDP-HD on highly overlapped and large size shapes datasets. (a) Clusters of D31 formed using CFSFDP. (b) Shows clusters formed by CFSFDP-HD. (c) Depicts the 15 clusters of R15 separated with CFSFDP. (d) Shows clusters R15 clusters created by CFSFDP-HD.

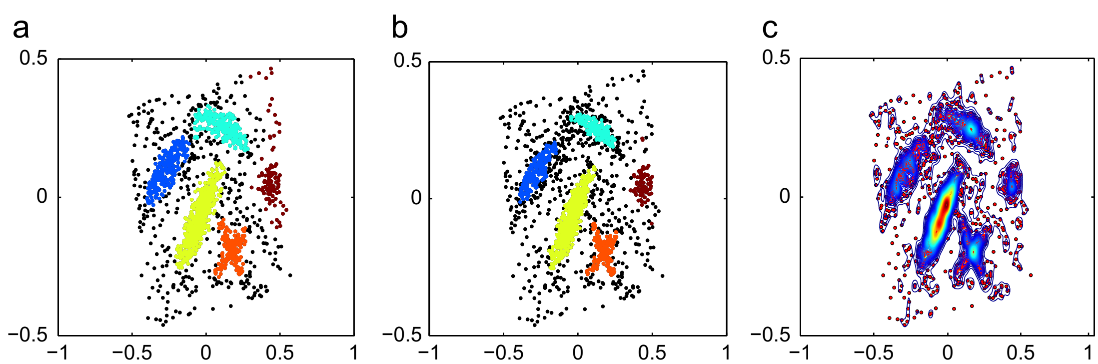

Fig. 4. Comparison of CFSFDP and CFSFDP-HD on the dataset generated by a probability distribution with non-spherical and strongly overlapping peaks. (a) Shows the clusters formed by CFSFDP that includes most of noise points as border points, the red cluster visualize this fact more clearly. (b) Shows the clusters created by CFSFDP-HD, in which density is expressed in a better way. (c) Shows the visualization of densities estimated by heat-diffusion method. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

the Gaussian kernel,

$$\hat{f}(d;\, t) = \frac{1}{n}\sum_{j=1}^{n} \theta(d,\, d_j;\, t), \quad d \in [0,\, 1], \tag{10}$$

where the theta kernel is

$$\theta(d,\, d_j;\, t) = \sum_{i=-\infty}^{\infty} \varnothing(d,\, 2k + d_i;\, t) + \varnothing(d,\, 2k - d_i;\, t) \tag{11}$$

Then Eq. (10) can be written as

$$\hat{f}(d;\, t) = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=-\infty}^{\infty} e^{-k^2\pi^2 t/2}\cos(k\pi d)\cos(k\pi d_j), \tag{12}$$

and Eq. (12) can be approximated as

$$\hat{f}(d;\, t) \approx \sum_{k=0}^{n-1} a_k e^{-k^2\pi^2 t/2}\cos(k\pi d), \tag{13}$$

where n is a large positive integer, and a_k is

$$a_k = \begin{cases} 1 & k = 0 \\ \frac{1}{n}\sum_{i=1}^{n}\cos(k\pi d_i) & k = 1,\, 2,\, \ldots,\, n-1 \end{cases}$$

Eq. (13) is a fully adaptive and alternative form of KDE and considers both the optimal bandwidth selection and the boundary corrections. Furthermore, Eq. (13) can be solved using fast Fourier transform and takes $O(n\log_2 n)$ operations [31,33]. For small bandwidth, Eq. (13) behaves like a Gaussian kernel and for large bandwidth like a uniform kernel [31,32]. It provides better performance and is consistent with the true density, whereas Eq. (3) is inconsistent [31,33]. The superior performance and fast evaluation of KDEs via diffusion is discussed in [32].

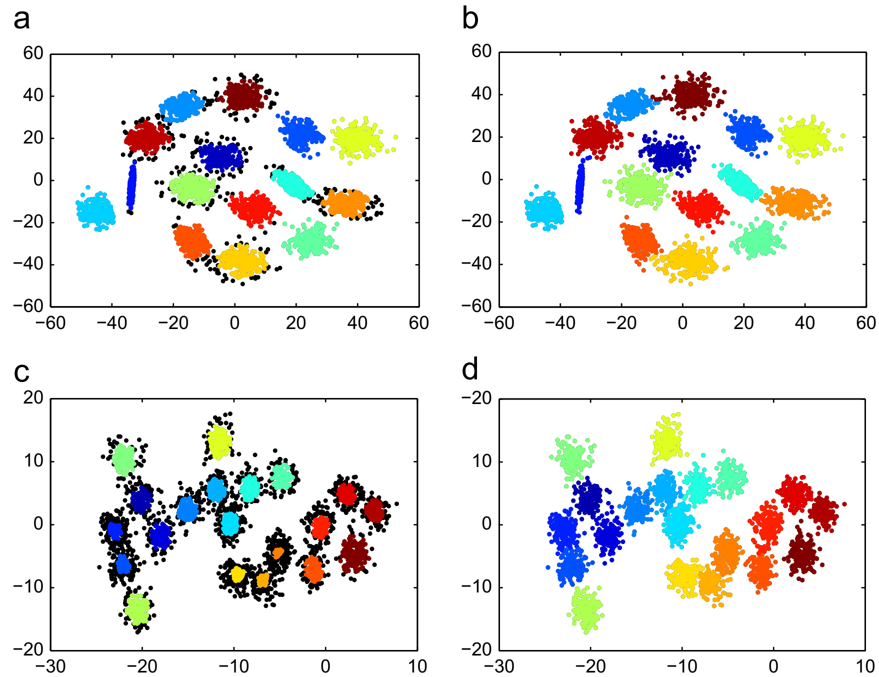

Fig. 5. Comparison of propose method with CFSFDP over large size datasets. (a) S1 dataset, clustered by CFSFDP, which identify most of boarder points as noise. (b) Shows clusters of S1 created by CFSFDP-HD. (c) CFSFDP also creates poor clusters on A1 dataset. It misclassify border points as noise at each cluster border region. (d) Clusters of A1 dataset, formed by CFSFDP-HD, which validates its effectiveness.

Table 2
Comparison of CFSFDP and proposed CFSFDP-HD to detect the core-points and misclassification of clusters points.

Dataset	CFSFDP		CFSFDP-HD	
	(identified core points)	(misclassified points)	(identified core points)	(misclassified points)
Aggregation	703	85	778	0
flame	83	157	240	0
Path-based spiral	312	0	312	0
R15	590	10	600	0
D31	1080	2020	2966	0
Dim2	1351	0	1351	0
toys problem	108	192	300	0
A1	1737	1263	3000	0
Diamond	2067	933	2666	0
S1	4756	244	5000	0

3.3. Optimal bandwidth selection

The improved Sheather–Jones (ISJ) [31] algorithm is used to calculate the optimal bandwidth. Optimal bandwidth is obtained as a fixed-point solution to a recursion and can be estimated using the fast cosine transform without considering the normality assumption on the distribution [31–33]. Botev et al. [31] proposed a unique solution of the nonlinear equation to adaptively find the optimal bandwidth t for KDE,

$$t = \xi\gamma^{[l]}\left(t\right). \tag{14}$$

The detailed description of the ISJ method is provided in [31].

The optimal bandwidth, t, scales the kernel function to estimate more accurate densities. We use $t = sqrt(t)/3.3$; to refine the border points of the clusters.

Algorithm 1. Clustering by fast search and find of density peaks via heat-diffusion.

Require: D distance matrix of dataset
Output: Organized clusters
1. Calculate t from Eq. (14)
2. Calculate ρ_i for point i from Eq. (13)
3. Calculate δ_i for point i from Eq. (2)
4. Plot ρ and δ on decision graph
5. Select cluster centers from the decision graph
6. Assign remaining points to cluster centers
7. Check the border point conditions for created clusters

4. Experiments

To evaluate the robustness of the proposed CFSFDP-HD method, we compared results from our proposed method to standard CFSFDP, AP, mean shift, and K-means methods on synthetic clustering datasets.

4.1. Datasets

We used 11 synthetic benchmark datasets, as shown in Table 1.

4.2. Results and discussion

To evaluate the performance of our proposed CFSFDP-HD method, we utilized the toys problem dataset and compared the identified clusters with that of CFSFDP. Fig. 2(a) shows the real clusters of the toys problem dataset. Fig. 2(b) and (c) shows the clusters formed by considering d_c as 2% and 1% of the entire dataset, respectively, as suggested from conventional CFSFDP. Even after tuning d_c, CFSFDP still identifies most of the core data points as noise. Fig. 2(d) shows the clusters from our proposed method, and Fig. 2(e) is the visualization of the density estimated by CFSFDP-HD. CFSFDP-HD expresses the true density potential of each object in the dataset and provides a strong foundation for

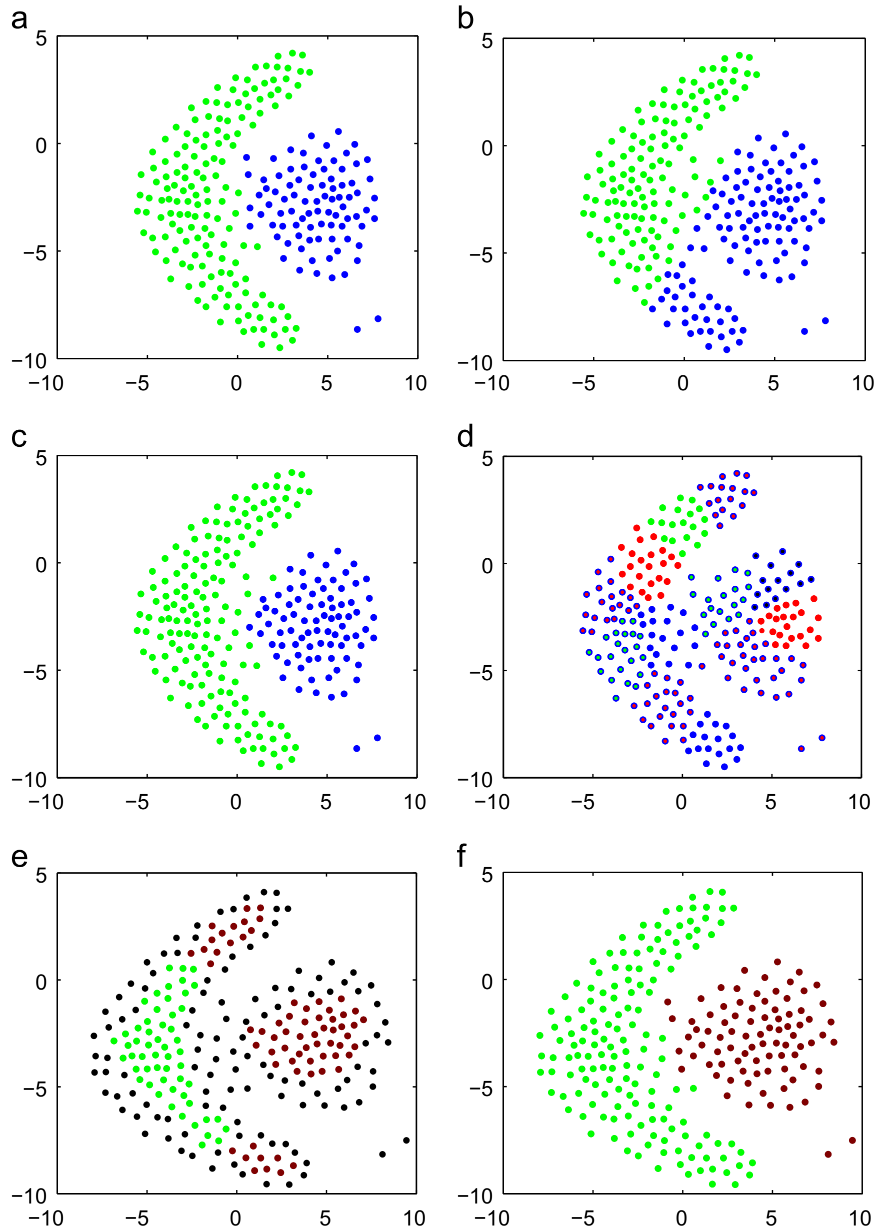

Fig. 6. Comparison of proposed CFSFDP-HD with different famous clustering algorithms. (a) Shows the ideally separated 2 clusters of flame dataset. (b) Shows the two clusters obtained by K-means clustering at $k=2$. (c) Two clusters created by mean shift clustering method at optimal size of window. (d) 13 clusters created by Affinity Propagation clustering of flame dataset. (e) Two clusters created by using CFSFDP in flame dataset. (f) Ideal clusters created by CFSFDP-HD of flame dataset.

identifying the number of expected clusters and border points. The time parameter of the heat equation, t, is an effective tool to find the border point densities.

Fig. 3(a) shows the 31 clusters of the D31 dataset as defined from CFSFDP. However, most of the cluster core points are identified as noise, and it cannot create the clusters effectively, even on the larger datasets. The density estimation and selection method using d_c misleads the creation of better clusters. The proposed CFSFDP-HD method correctly identifies the clusters and successfully separates border points in the dataset with overlapping densities, as shown in Fig. 3(b). The comparison of CFSFDP and CFSFDP-HD on the R15 dataset is shown in Fig. 3(c) and (d), respectively. Clusters created by CFSFDP-HD are more consistent and effective than those from CFSFDP. CFSFDP identified 10 core points as noise, whereas CFSFDP-HD created accurate clusters without any misclassification of data points.

Fig. 4 compares CFSFDP-HD and CFSFDP on the point distribution dataset generated and utilized by [30]. This dataset contains arbitrary shaped clusters with high overlapping density peaks and noise. Fig. 4(a) shows CFSFDP clusters include some of the noise at border regions as core points (see the red cluster in Fig. 4(a) particularly). Fig. 4(b) shows CFSFDP-HD clusters with better performance, and demonstrates the ability to separate the noise from cluster border points using t rather than d_c. Fig. 4 (c) shows a 3D view of the estimated density of each data point using our proposed method.

The effectiveness of CFSFDP-HD over CFSFDP on different datasets (S1, D1, aggregation, diamond, path-based spiral, and dim2) was also analyzed. These results are not shown here in detail due to space limitations, but are summarized in Fig. 5. In general, CFSFDP could not express the compact relationship of densities at border points of the clusters, and clusters created by CFSFDP-HD are more accurate. The compact clusters created by CFSFDP-HD validate the models robustness over large datasets. However, CFSFDP-HD is equally effective for smaller datasets. The detailed comparison is given in Table 2.

Fig. 6 shows the detailed comparison of CFSFDP-HD with current state-of-the-art methods—K-means (Fig. 6(b)) [50], mean shift (Fig. 6(c)) [51], AP clustering (Fig. 6(d)) [52], and CFSFDP (Fig. 6(e))—using the flame synthetic dataset. The real clusters of the data set are shown in Fig. 6(a).

For $k=2$, K-means could not find the relation between connected densities. Hence, K-means is not suitable for clustering the dataset that follows some distribution and have clusters of arbitrary shapes.

The accuracy and shape of clusters created by mean shift depends greatly upon the window size, which is hard to estimate. The optimal clusters of mean shift were obtained at window$=3.1$. Even at this optimal window size, the mean shift method misclassifies four points.

The 13 clusters of the flame dataset created by AP (Fig. 6(d)) are very different from those of Fig. 6(a).

The clusters created by CFSFDP (Fig. 6(e)) were for $d_c=0.71$, which is 1% of the entire dataset. CFSFDP identified most of the cluster core data points as noise and also could not find a compact relation of connected densities. We attempted to tune d_c, but could not improve on these results.

The proposed CFSFDP-HD method (Fig. 6(f)) successfully separated the flame dataset into two clusters, and these clusters are similar to the real clusters of Fig. 6(a).

Compared to the tested methods, the proposed CFSFDP-HD is more robust and effective. K-means and AP are partition based clustering methods. Both methods partition the data into the spherical shapes of clusters. Hence, could not find arbitrary cluster shapes. The accuracy of mean shift and CFSFDP depends upon tunable parameters, which are hard to estimate. In mean shift, the size and shape of clusters is subject to window size. However, in CFSFDP, the cutoff distance is hard to estimate and it also uses different methods to estimate densities depending upon the nature of the given data. The CFSFDP-HD is capable to find arbitrary shapes of clusters and uses adaptive way to estimate densities, efficiently and effectively. Hence, the clustering results of CFSFDP-HD are more consistent and robust as compare to K-means, AP, mean shift, and CFSFDP.

Table 2 shows the detailed comparison between CFSFDP and CFSFDP-HD in terms of identified cluster points and misclassified cluster points. The proposed CFSFDP-HD method identifies cluster core points more accurately independent of the nature of dataset, whereas CFSFDP ability for finding the densities and border points depends highly on the nature of the dataset. Thus the proposed CFSFDP-HD method is an effective generalized solution to cluster different datasets.

5. Conclusion

A new method, CFSFDP-HD, based on the heat equation was proposed to better estimate the densities for creating more accurate clusters and to more effectively separate noise from clusters points. Based on the heat diffusion equation on an infinite domain, the proposed method incorporates the d_c cutoff distance selection and the boundary correction of KDE. Therefore, the overhead involved in estimating the densities of data points more accurately and the selection of the sensitive cutoff parameter in CFSFDP are removed. Tests conducted on 11 synthetic datasets showed the robustness and effectiveness of the proposed method compared to CFSFDP and other current state-of-the-art methods.

In CFSFDP and CFSFDP-HD, decision graph is used to select cluster centers with human interaction. Human based selection of cluster centers is a potential barrier in automatic analysis of data. In the future work, we will try to extend CFSFDP-HD to a fully adaptive method.

Acknowledgments

This research is sponsored by National Natural Science Foundation of China (Nos. 61171014, 61371185, 61401029, 61472044, 61472403, 61571049) and the Fundamental Research Funds for the Central Universities (Nos. 2014KJJCB32, 2013NT57) and by SRF for ROCS, SEM.

References

[1] Lana Yeganova, Won Kim, Sun Kim, W.J. Wilbur, Retro: concept-based clustering of biomedical topical sets, Bioinformatics 30 (22) (2014) 3240–3248.
[2] Chen Xu, Zhengchang Su, Identification of cell types from single-cell transcriptomes using a novel clustering method, Bioinformatics 37 (10) (2015) 2041–2256.
[3] Suzuki Shuji, Masanori Kakuta, Takashi Ishida, Yutaka Akiyama, Faster sequence homology searches by clustering subsequences, Bioinformatics 31 (8) (2015) 1183–1190.
[4] Y. Yan, Y. Qian, H. Sharif, D. Tipper, A survey on cyber security for smart grid communications, IEEE Commun. Surv. Tutor. 14 (4) (2012) 998–1010.
[5] Portnoy, Leonid, Eleazar Eskin, Sal Stolfo, Intrusion detection with unlabeled data using clustering, in: Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA-2001), 2001, pp. 5–8.
[6] Jiwen Lu, Venice Erin Liong, Xiuzhuang Zhou, Jie Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 37 (10) (2015) 2041–2256.
[7] Jiwen Lu, Xiuzhuang Zhou, Yap-Peng Tan, Yuanyuan Shang, Jie Zhou, Neighborhood repulsed metric learning for kinship verification, IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI) 36 (2) (2014) 331–345.
[8] Jiwen Lu, Yap-Peng Tan, Gang Wang, Discriminative multimanifold analysis for face recognition from a single training sample per person, IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI) 35 (1) (2013) 39–51.
[9] Jiwen Lu, Venice Erin Liong, Jie Zhou, Cost-sensitive local binary feature learning for facial age estimation, IEEE Trans. Image Process. (T-IP) 24 (12) (2015) 5356–5368.
[10] Jiwen Lu, Gang Wang, Weihong Deng, Kui Jia, Reconstruction-based metric learning for unconstrained face verification, IEEE Trans. Inf. Forensics Secur. (T-IFS) 10 (1) (2015) 79–89.
[11] Jiwen Lu, Yap-Peng Tan, Gang Wang, Gao Yang, Image-to-set face recognition using locality repulsion projections and sparse reconstruction-based similarity measure, IEEE Trans. Circuits Syst. Video Technol. (T-CSVT) 23 (6) (2013) 1070–1080.
[12] Jones, M. Kristen, M. Lacy, Measuring the clustering around normal and dust-obscured quasars at 2 in the Spitzer extragalactic representative volume survey (SERVS), in: American Astronomical Society Meeting Abstracts, vol. 223, no. 223, 2014.
[13] Chakrabarti, Soumen, Data mining for hypertext: a tutorial survey, in: ACM SIGKDD Explorations Newsletter, vol. 1, no. 2, 2000, pp. 1–11.
[14] Maw-Shang Chang, Li-Hsuan Chen, Ling-Ju Hung, Peter Rossmanith, Guan-Han Wu, Exact algorithms for problems related to the densest k-set problem, Inf. Process. Lett. 114 (9) (2014) 510–513.
[15] P. Lovely Sharma, K.A. Ramya, Review on density based clustering algorithms for very large datasets, Int. J. Emerg. Technol. Adv. Eng. 3 (12) (2013) 12.
[16] Glory H. Shah, C.K. Bhensdadia, Amit P. Ganatra, An empirical evaluation of density-based clustering techniques, Int. J. Soft Comput. Eng. (IJSCE) (2012) 2231–2307.
[17] M. Parimala, Daphne Lopez, N.C. Senthilkumar, A survey on density based clustering algorithms for mining large spatial databases, Int. J. Adv. Sci. Technol. 31 (1) (2011).
[18] Ester, Martin, Hans-Peter Kriegel, J. Sander, Xiaowei Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Kdd, vol. 96, no. 34, 1996, pp. 226–231.
[19] Derya Birant, S.T. Alp Kut, DBSCAN: an algorithm for clustering spatial-temporal data, Data Knowl. Eng. 60 (1) (2007) 208–221.
[20] Liu, Peng, Dong Zhou, Naijun Wu, VDBSCAN: varied density based spatial clustering of applications with noise, in: 2007 Service Systems and Service Management, 2007, pp. 1–4.
[21] Fionn Murtagh, Pedro Contreras, Algorithms for hierarchical clustering: an overview, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2 (1) (2012) 86–97.
[22] Chen, Na, Ze-shui Xu, Mei-mei Xia, Hierarchical hesitant fuzzy K-means clustering algorithm, Appl. Math.-A J. Chin. Univ. 29 (1) (2014) 1–17.
[23] Daniel, Jaeger, Johannes Barth, Anna Niehues, Christian Fufezan, pyGCluster, a novel hierarchical clustering approach, Bioinformatics 30 (6) (2014) 896–898.
[24] Julien Jacques, Cristian Preda, Functional data clustering: a survey, Adv. Data Anal. Classif. 8 (3) (2014) 231–255.
[25] Mukhopadhyay Anirban, Ujjwal Maulik, Sanghamitra Bandyopadhyay,

A. Carlos, Coello Coello, A survey of multiobjective evolutionary algorithms for data mining: part II, IEEE Trans. Evolut. Comput. 18 (1) (2014) 20–35.
[26] Kannuri Lahari, M. Ramakrishna Murty, Suresh C. Satapathy, Partition based clustering using genetic algorithm and teaching learning based optimization: performance analysis, Adv. Intell. Syst. Comput. 338 (2015) 191–200.
[27] Tao Chen, Nevin L. Zhang, Tengfei Liu, Kin Man Poon, Yi Wang, Model-based multidimensional clustering of categorical data, Artif. Intell. 176 (1) (2012) 2246–2269.
[28] Amandeep Kaur Mann, Navneet Kaur, Survey paper on clustering techniques, Int. J. Sci. Eng. Technol. Res. (IJSETR) 2 (4) (2013).
[29] Monali Parikh, Tanvi Varma, Survey on different grid based clustering algorithms, Int. J. Adv. Res. Comput. Sci. Manag. Stud. 2 (2) (2014).
[30] Alex Rodriguez, Alessandro Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.
[31] Z.I. Botev, J.F. Grotowski, D.P. Kroese, Kernel density estimation via diffusion, Ann. Stat. 38 (5) (2010) 2916–2957.
[32] Smita Krishnaswamy, Matthew H. Spitzer, Michael Mingueneau, Sean C. Bendall, Oren Litvin, Erica Stone, Dana Pe'er, Garry P. Nolan, Conditional density-based analysis of T cell signaling in single-cell data, Science 346 (6213) (2014) 1250689.
[33] Xiaoyuan Xu, Zheng Yan, Shaolun Xu, Estimating wind speed probability distribution by diffusion-based kernel density method, Electr. Power Syst. Res. 121 (2015) 28–37.
[34] K. Fukunaga, L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, IEEE Trans. Inf. Theory 21 (1975) 32–40.
[35] Y. Cheng, Mean shift, mode seeking, and clustering, IEEE Trans. Pattern Anal. Mach. Intell. 17 (8) (1995) 790–799.
[36] Eric L. Lehmann, Model specification: the views of Fisher and Neyman, and later developments, Stat. Sci. 5 (2) (1990) 160–168.
[37] M. Rosenblatt, Remarks on some nonparametric estimates of a density-function, Ann. Math. Stat. 27 (3) (1956) 832–837.
[38] Adam A. Margolin, Kai Wang, Wei Keat Lim, Manjunath Kustagi, Ilya Nemenman, Andrea Califano, Reverse engineering cellular networks, Nat. Protoc. 1 (2) (2006) 662–671.
[39] M. Chris Jones, James S. Marron, Simon J. Sheather, A brief survey of bandwidth selection for density estimation, J. Am. Stat. Assoc. 91 (433) (1996) 401–407.
[40] Simon J. Sheather, Michael C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, J. R. Stat. Soc. Ser. B-Methodol. 53 (3) (1991) 683–690.
[41] A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation, ACM Trans. Knowl. Discov. Data (TKDD) 1 (1) (2007) 1–30.
[42] L. Fu, E. Medico, FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data, BMC Bioinform. 8 (3) (2007).
[43] H. Chang, D.Y. Yeung, Robust path-based spectral clustering, Pattern Recognit. 41 (2) (2008) 191–203.
[44] C.J. Veenman, M.J.T. Reinders, E. Backer, A maximum variance cluster algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 24 (9) (2002) 1273–1280.
[45] P. Franti, O. Virmajoki, V. Hautamaki, Fast agglomerative clustering using a k-nearest neighbor graph, IEEE Trans. Pattern Anal. Mach. Intell. 28 (11) (2006) 1875–1881.
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[47] Karkkainen, Ismo, Pasi Franti, Dynamic local search for clustering with unknown number of clusters, in: Proceedings of International Conference on Pattern Recognition, vol. 16, no. 2, 2002, pp: 240–243.
[48] Salvador, Stan, Philip Chan, Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms, in: Proceedings of International Conference on Tools with Artificial Intelligence, ICTAI, 2004, pp. 576–584.
[49] Pasi Franti, Olli Virmajoki, Iterative shrinking method for clustering problems, Pattern Recognit. 39 (5) (2006) 761–775.
[50] James MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, no. 14, 1967, pp. 281–297.
[51] Dorin Comaniciu, Peter Meer, Mean shift: a robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 603–619.
[52] Brendan J. Frey, Delbert Dueck, Clustering by passing messages between data points, Science 315 (5814) (2007) 972–976.

Rashid Mehmood has received his Masters degree from COMSITS, Islamabad, Pakistan 2012. He is currently working toward the Ph.D. degree in the School of Information Sciences and Technology at Beijing Normal University, Beijing, China. His current research interest includes clustering, DNA barcode analysis.

Guangzhi Zhang is currently a student in Beijing Normal University, where he is making efforts to obtain a Ph.D. degree. He devotes himself to the research of big data and the internet of things, especially medical big data.

Rongfang Bie is currently a Professor at the College of Information Science and Technology of the Beijing Normal University. She received her M.S. degree on June 1993 and Ph.D degree on June 1996 from Beijing Normal University. She was with the Computer Laboratory at the University of Cambridge as a visiting faculty from March 2003 for one year. She is the author or co-author of more than 80 papers. Her current research interests include knowledge representation and acquisition for the Internet of Things, database application and data mining, software reliability engineering and model theory, and cognitive radio networks.

Hassan Dawood has received his Master's and Doctorate of Engineering degree in Computer Application Technology from Beijing Normal University, Beijing; in 2012 and 2015, respectively. He has published several well-known journals and conferences papers. His current research interest includes image processing, pattern recognition, and feature extraction.

Haseeb Ahmad received the B.S. degree in Mathematics from G.C. University, Faisalabad, Pakistan in 2010 and the Master's degree in Computer Science from Virtual University of Pakistan in 2012. He is currently a Ph.D. student in School of Computer Science at Beijing University of Posts and Telecommunications, Beijing, China. His current research interest includes information security, clustering and image processing.