IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 10, 2019, accepted September 1, 2019, date of publication September 4, 2019,
date of current version September 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939437

Clustering by Search in Descending Order and Automatic Find of Density Peaks

TONG LIU, HANGYU LI, AND XUDONG ZHAO
College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China

Corresponding author: Xudong Zhao (zhaoxudong@nefu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2572018BH01, in part by the National Undergraduate Innovation Project under Grant KY2017032 and Grant 201910225184, and in part by the Specialized Personnel Start-up Grant (Also National Construction Plan of World-class Universities and First-class Disciplines) under Grant 41113237.

ABSTRACT *Clustering by fast search and find of density peaks* published on journal *Science* in 2014 is a density-based clustering technique, which is not only unnecessary to determine the number of clusters in advance, but also able to recognize the clusters of arbitrary shapes. Due to a manual selection of clustering centers on a decision graph, samples which belong to one cluster may be assigned to two or more clusters and vice versa. On assumption that boundary points which keep comparable densities with cluster centers should be regarded as inner points, we make a new method which not only can find all possible clusters automatically but also can combine those with similarities simultaneously to obtain the final clusters. Unlike clustering by fast search and find of density peaks, we only focus on densities with discarding the relative metric which measures the minimum distance between a cluster center and a point with a higher density. Qualitative and quantitative experimental results on sufficient datasets demonstrate the effectiveness of our method.

INDEX TERMS Density-based clustering, density peaks clustering, automatic clustering, density categorization, cluster merging.

I. INTRODUCTION

Clustering by fast search and find of density peaks [1], which can discover arbitrary shapes regardless of determining the number of clusters in advance, has yielded good performance in many fields. In contrast with density-based clustering such as DBSCAN [2], DBCLASD [3] and DENCLUE [4], *etc*., this method asserts no need of choosing any threshold in advance. However, cluster centers have to be interactively appointed on a decision graph. When cluster centers are manually selected, samples derived from an original cluster might be inappropriately divided and vice versa. In fact, this phenomenon can be shown in Fig. 1(a). Intuitively, one may choose one to six points to be cluster centers on the decision graph.

Many approaches have been proposed for further improvements. Several of prevailing literatures are listed as follows. Density-ratio is proposed for discovering clusters with varying densities [5]. A nonparametric method is presented for both selection of the cutoff distance and boundary correction of the kernel density estimation via heat diffusion [6]. Besides, a fuzzy clustering is proposed for adaptively selecting the cluster centers effectively [7]. A comparison whether the minimum distance between a cluster center and a point with a higher density is large or not is made [8]. In order to calculate the similarity between points, shared neighbor information is additionally invoked [9]. In addition, a novel statistical outlier detection method is designed to identify cluster centers [10].

In fact, the clustering method named as fast search and find of density peaks (DPC) is based on two assumptions. Firstly, the cluster center within each cluster keeps a highest local density. That is to say, any sample with a density larger than a cluster center should belong to another cluster other than this one. Secondly, any distance between a cluster center and a sample which not only keeps a higher local density but also derives from a different cluster, is commonly large.

As a matter of fact, the application of the second assumption that metric δ measures a local separation by computing the minimum distance between a current point and another one with a higher density for finding cluster centers keeps a small flaw. Taking the samples of the cluster labeled

The associate editor coordinating the review of this article and approving it for publication was Qilian Liang.

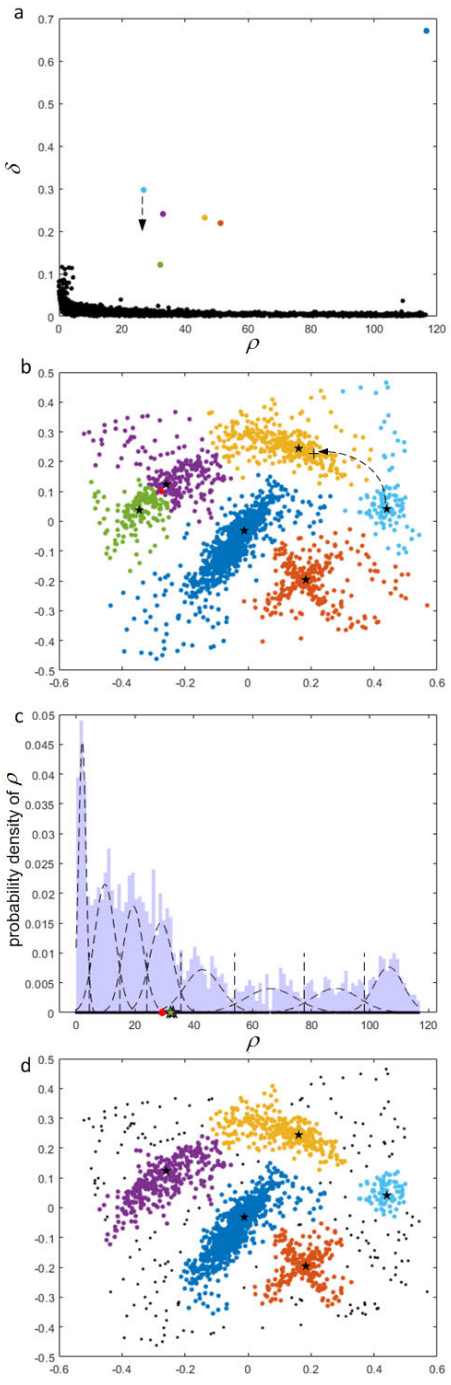


FIGURE 1. Comparative results of clustering by fast search and find of density peaks and our method using a synthetic dataset. (a) A decision graph derived from clustering by fast search and find of density peaks. (b) The clustering result using fast search and find of density peaks. Here, icon ★ represents the cluster center. Icon • labeled red denotes a boundary point with a relatively high density. (c) A density-based decision graph using GMM-based model selection. (d) The experimental result after traversing and clustering all samples according to their densities in descending order and merging clusters with similarities.

wathet in Fig. 1(b) as an example, the metric δ of its cluster center corresponding to the wathet point in Fig. 1(a) becomes smaller when the samples of the cluster labeled wathet in Fig. 1(b) move closer to the point (i.e., marked as icon "+") not only with a higher density than the cluster center but also keeping a minimum distance to it. At last, the cluster labeled wathet may be wrongly merged into the cluster labeled ochre, due to an intuitively small δ on the decision graph. This trend can be seen along the dotted arrows, which are correspondingly shown in Fig. 1(a) and Fig. 1(b). Therefore, metric δ is to be discarded.

Focusing only on densities, we propose a new clustering method for automatic find of density peaks. In fact, automatic clustering derives from the fact that samples can be stratified into cluster centers, inner points and boundary points according to their densities. Firstly, we sort all the samples with their densities in descending order. Secondly, we categorize densities using GMM-based model selection for further merging step considering the one-dimensional characteristics of densities. Last but most important, we present an algorithm to traverse all the samples for automatically finding clusters as many as possible and combining those with similarities on assumption that boundary points which keep comparable densities with cluster centers should be considered to be inner points. Qualitative and quantitative experimental results on sufficient datasets distinctly demonstrate the effectiveness of our method.

II. BACKGROUND

We denote a dataset as $D = \{x_i | i = 1, 2, 3, \ldots, n\}$, where n represents the sample size. The local density ρ_i of sample i is defined as

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

where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise, and d_{ij} is a distance of sample i from sample j. d_c is a cutoff distance, which generally takes the first 2% of all distances based on experience. Here, the local density ρ_i represents the number of samples with their distances from sample i less than d_c. Gaussian density is also considered when data is rare, which is defined as

$$\rho_i = \sum_j e^{-(d_{ij}/d_c)^2}. \tag{2}$$

As to the invoked GMM-based model selection [11] which helps to stratify densities of samples, it refers to a penalized likelihood for estimating finite Gaussian mixture models and a modified EM algorithm to simultaneously select the number of components and estimate the unknown parameters. For the sake of simplicity, one can choose the python package sklearn.mixture.gaussian as alternative.

III. PROPOSED METHOD

When δ is discarded, what needs to be concerned is only ρ. Following the first assumption of DPC, a point with the maximum density must be a cluster center. Thus, what needs to be done first is to sort all the samples in descending order according to their densities. Besides, any point i with a high

density may be a cluster center or an inner point, but not a cluster boundary (i.e., its neighboring points j with $d_{ij} < d_c$ are assigned to different clusters). That's to say, if a point keeps a comparable density with an assigned cluster center, it cannot be a cluster boundary. This just provides a strategy to decide whether to merge clusters or not. Therefore, density categorization is to be made right before clustering of local densities.

A. SAMPLE REORDERING

We first calculate the local density ρ_i of each sample i using equation (2) by taking the first 2% of all Euclidean distances as the cutoff value d_c, considering the limited sample size of data. Sample j is viewed as a neighboring point of sample i, if $d_{ij} < d_c$. Then, all samples are sorted with their densities in descending order for further calculation, which is expressed as a density vector P where $P = (\rho_1, \rho_2, \ldots, \rho_n)$.

B. DENSITY CATEGORIZATION

Then, we categorize one-dimensional densities into groups using GMM-based model selection [11] and make a density-based decision graph, as shown in Fig. 1(c). Here, the transverse axis refers to the densities with their values labeled as black scatters attached to the axis. The cyan bars correspond to histograms. The dashed lines refer to probability densities of the obtained Gaussian distributions, which are expressed as a density Gaussian set G where $G = \{g_1, g_2, \ldots, g_k\}$. It can be seen in Fig. 1(c) that the boundary point labeled red keeps a comparable density with two cluster centers labeled purple and green, for their densities fall into a same Gauss. The obtained categories are used for further cluster merging step.

C. CLUSTERING OF LOCAL DENSITIES

We focus on ρ which is considered as a global calculation rather than δ. After calculating ρ_i of each sample i, sequencing them in descending order and fulfilling density categorization, we begin to traverse all the samples in that order. The corresponding algorithm is illustrated in Algorithm 1.

Here, $N(i)$ denotes the set of sample index corresponding to the neighboring points of current point i. Besides, $\lambda_{N(i)}$ refers to the corresponding labels. For each point, we judge whether its neighboring points with greater densities belong to any existing cluster. If its neighboring points with higher densities are all assigned to a same cluster, the current point must be also assigned to that cluster. If its neighboring points with greater densities are assigned to several clusters, it will need to be further judged whether there is a cluster which a neighboring point belongs to and its cluster center keeps the same Gaussian distribution of density as the current point does. That is, ρ_i and $\min \rho_{C_S} \in g_r$, where C_S is a subset of set C and serves as the set of sample index corresponding to the neighboring label set S. If so, clusters which the neighboring points belong to are to be merged into one. If not, the current point will be assigned to the cluster which keeps the labeled point with the minimum distance to the point being assigned.

Algorithm 1 Clustering After Sample Reordering and Density Categorization

Require:
Density vector $P = (\rho_1, \rho_2, \ldots, \rho_n)$, each component of which is sorted in descending order
Density Gaussian set $G = \{g_1, g_2, \ldots, g_k\}$
Ensure: Cluster center set C, label set $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$

initialize d_c, $C \Leftarrow \phi$, each $\lambda_i \Leftarrow 0$, $label \Leftarrow 0$
for $i \Leftarrow 1$ to n **do**
  if $\exists\ \lambda_{N(i)} \neq 0$ **then**
    $S \Leftarrow \{\lambda_{N(i)}\} \setminus \{0\}$ //(n-1) *step*
    if $|S| = 1$ **then**
      $\lambda_i \Leftarrow$ the element of S // 1 *step*
    else
      if ρ_i and $\min \rho_{C_S} \in g_r$, where $r = 1, \ldots, k$ **then**
        //*merging step*
        $\lambda_l, \lambda_i \Leftarrow \lambda_{\arg\max(\rho_{C_S})}, \forall\ \lambda_l \in S$ // n *step*
        $C \Leftarrow C \setminus C_s$ // 1 *step*
        $C \Leftarrow C \bigcup \{\arg\max(\rho_{C_s})\}$ // n *step*
      else
        $\lambda_i \Leftarrow \lambda_{\arg\min_j(d_{ij})}$, where $\lambda_j \neq 0$ // n *step*
      end if //Condition expression: $(n * logn)$ *step*
    end if //Condition expression: 1 *step*
  else
    if ρ_i is not in the last Gauss $\wedge \rho_i \neq 0$ **then**
      $C \Leftarrow C \cup \{i\}$ // 1 *step*
      $label \Leftarrow label + 1, \lambda_i \Leftarrow label$ // 1 *step*
    end if //Condition expression: 1 *step*
  end if //Condition expression: (n-1) *step*
end for //Loop: (n+1) *step*, $O(n^2 logn)$

If its neighboring points don't belong to any existing cluster, it will also need to be further judged whether the point being assigned has fallen into the minimum Gaussian distribution of density or its density is zero. If so, it will be regarded as noise with its label $\lambda_i = 0$. Otherwise, it may be viewed as a new cluster center. In that case, we'll make a new cluster center label and put the sample index into the cluster center set C.

IV. EXPERIMENTAL RESULTS

In order to show the effectiveness of our method, we generate a synthetic dataset illustrated in Fig. 1. In turn, the decision graph and the clustering result using DPC [1] are shown in Fig. 1(a) and Fig. 1(b). A density-based decision graph and an automatic clustering result using our approach are listed in Fig. 1(c) and Fig. 1(d), respectively. It can be apparently seen that the proposed automatic clustering method keeps a better result than clustering using DPC.

In addition, we select 15 groups of artificial data to test the effectiveness of our algorithm by comparing with DPC [1], DBSCAN [2], DPC via heat diffusion (DPC-HD) [6] and fuzzy clustering by DPC (Fuzzy-DPC) [7]. We make

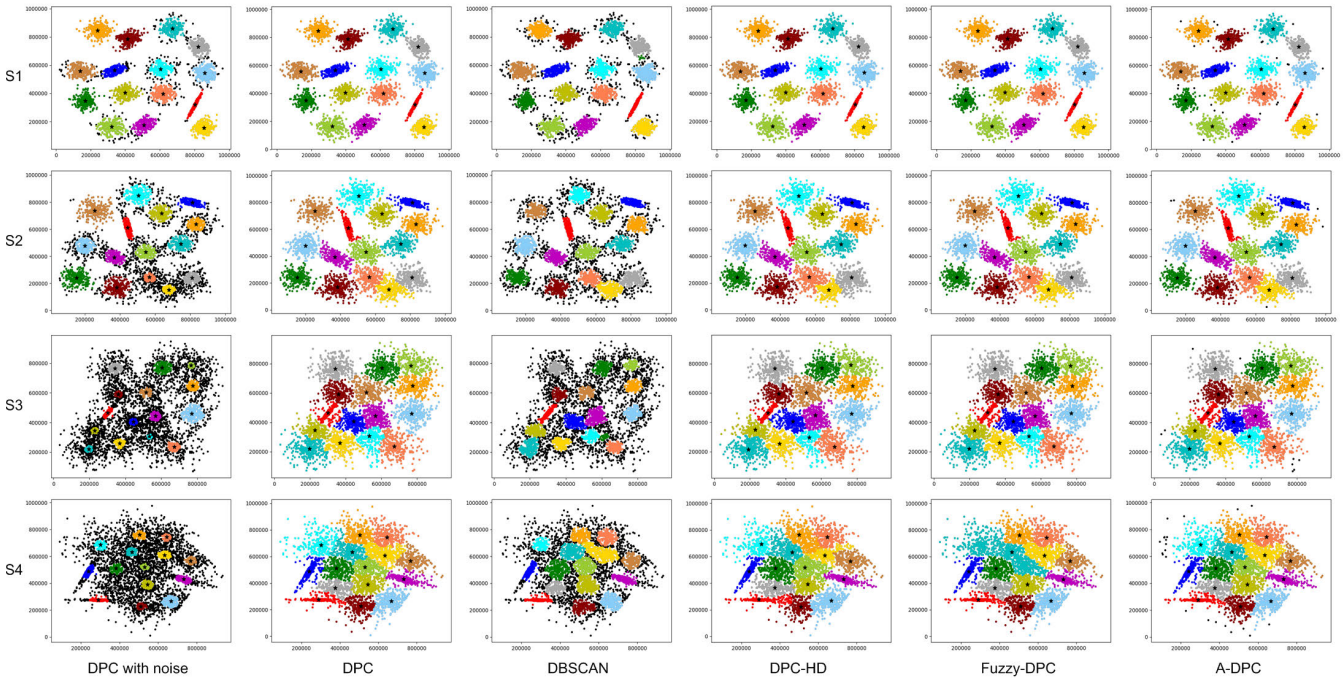


FIGURE 2. Qualitative results of clustering on data S1, S2, S3 and S4 [12].

TABLE 1. Respective optimal parameters on each data.

Data	DPC + noise	DPC	DBSCAN	DPC-HD	Fuzzy-DPC	A-DPC
S1	d_c=0.02	d_c=0.02	ϵ=0.015,$MinPts$=8	–	d_c=0.02	d_c=0.02
S2	d_c=0.02	d_c=0.02	ϵ=0.02,$MinPts$=20	–	d_c=0.02	d_c=0.02
S3	d_c=0.02	d_c=0.02	ϵ=0.018,$MinPts$=22	–	d_c=0.02	d_c=0.02
S4	d_c=0.02	d_c=0.02	ϵ=0.025,$MinPts$=40	–	d_c=0.02	d_c=0.02
Aggregation	d_c=0.02	d_c=0.02	ϵ=0.0526,$MinPts$=14	–	d_c=0.02	d_c=0.02
Compound	d_c=0.024	d_c=0.024	ϵ=0.042,$MinPts$=6	–	d_c=0.024	d_c=0.024
Flame	d_c=0.08	d_c=0.08	ϵ=0.057,$MinPts$=4	–	d_c=0.08	d_c=0.08
D31	d_c=0.01	d_c=0.01	ϵ=0.016,$MinPts$=7	–	d_c=0.008	d_c=0.01
R15	d_c=0.03	d_c=0.03	ϵ=0.028,$MinPts$=12	–	d_c=0.01	d_c=0.03
Spiral	d_c=0.04	d_c=0.04	ϵ=0.06,$MinPts$=2	–	d_c=0.04	d_c=0.04
Pathbased	d_c=0.02	d_c=0.02	ϵ=0.68,$MinPts$=10	–	d_c=0.02	d_c=0.02
Jain	d_c=0.032	d_c=0.032	ϵ=0.058,$MinPts$=2	–	d_c=0.032	d_c=0.032
Revised pathbased	d_c=0.0156	d_c=0.0156	ϵ=0.07,$MinPts$=6	–	d_c=0.0156	d_c=0.0156
Revised jain	d_c=0.042	d_c=0.042	ϵ=0.07,$MinPts$=6	–	d_c=0.032	d_c=0.032
T48k	d_c=0.012	d_c=0.012	ϵ=0.022,$MinPts$=16	–	d_c=0.012	d_c=0.012
T58k	d_c=0.035	d_c=0.035	ϵ=0.03,$MinPts$=20	–	d_c=0.046	d_c=0.035
T170k	d_c=0.01	d_c=0.01	ϵ=0.022,$MinPts$=16	–	d_c=0.01	d_c=0.01
Olivetti Face Database	–	–	–	–	–	d_c=0.0247

comparisons with the four algorithms because of their similarities in methods. DPC utilizes the order of local densities. As to DBSCAN, it considers density neighborhood for assignment. On each data, we use the iterative way, select the respective optimal parameters, and perform multiple clustering operations to list the optimal experimental results. Corresponding parameters are listed in Table 1. Noise is considered to be in existence by default. Anyway, noise is also thought to be not existent in order to show better results using pure DPC (i.e., ''DPC'' rather than ''DPC with noise'' or ''DPC + noise''). As to GMM model selection before A-DPC, the initial number of components is 15. K-means is used for calculation of the initial weights. The convergence threshold and the maximum number of iteration are 0.001 and 100, respectively.

Firstly, we select a dataset containing four data [12] (i.e., S1, S2, S3 and S4) for making qualitative and quantitative comparisons among DPC [1], DBSCAN [2], DPC-HD [6], Fuzzy-DPC [7] and our automatic DPC (A-DPC). Considering that all the four data are of no original labels, we select three internal criteria, i.e., Davies-Bouldin score (DBI) [13], Silhouette coefficient (SC) [14] and

TABLE 2. Quantitative results of clustering on data S1, S2, S3 and S4 [12].

Data	DBI [13]			SC [14]			CH [15]		
S1	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	1.4129	**0.3662**	1.4406	0.6681	**0.7110**	0.4696	5997.9319	**22608.5925**	1418.2341
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	0.3663	**0.3662**	1.5525	0.7109	**0.7110**	0.7078	22603.3956	**22608.5925**	15815.6320
S2	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	1.6675	0.4706	1.9644	0.3955	**0.6220**	0.2953	1204.6509	**13249.7742**	653.0391
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	0.4698	0.4706	2.5703	0.6194	**0.6220**	0.6210	13114.6275	**13249.7742**	11270.1868
S3	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	1.5534	**0.6561**	1.6343	-0.1842	**0.4851**	-0.0679	171.8577	**7641.4448**	223.0208
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	0.6632	**0.6561**	3.3972	0.4762	**0.4851**	0.4843	7353.9319	**7641.4448**	6536.6902
S4	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	2.9544	**0.6817**	3.3236	-0.1000	**0.4654**	0.0776	251.6745	**5852.8935**	362.9914
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	0.7549	0.8218	2.5181	0.4320	0.3917	0.4642	5245.3659	3455.0292	4710.0661

TABLE 3. Quantitative results of clustering on six data with original labels.

Data	MI [21]			ARI [22]			FMI [23]		
Aggregation	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	1.5136	**1.6876**	1.6778	0.8353	**0.9978**	0.9888	0.8705	**0.9983**	0.9912
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	1.6042	**1.6876**	**1.6876**	0.9063	**0.9978**	**0.9978**	0.9265	**0.9983**	**0.9983**
Compound	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	1.1876	1.1876	1.3180	0.7826	0.7826	0.8795	0.8548	0.8548	0.9118
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	1.1876	1.2064	**1.4712**	0.7826	0.5895	**0.9658**	0.8548	0.6915	**0.9743**
Flame	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	0.2221	**0.6548**	0.6444	0.1615	**1.000**	0.9659	0.6136	**1.0000**	0.9840
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	0.6548	**0.6548**	**0.6548**	**1.0000**	**1.0000**	0.9881	**1.0000**	**1.0000**	0.9945
D31	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	1.9063	**3.2893**	2.7072	0.1011	**0.9370**	0.4311	0.2308	**0.9390**	0.4776
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	3.2842	3.2884	3.2877	0.9346	0.9364	0.9352	0.9366	0.9385	0.9373
R15	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	2.6168	**2.6924**	2.4949	0.9317	**0.9928**	0.8656	0.9364	**0.9932**	0.8745
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	2.6924	2.6741	2.6800	**0.9928**	0.9821	0.9817	**0.9932**	0.9833	0.9829
Spiral	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	1.0984	**1.0984**	**1.0984**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	1.0984	**1.0984**	**1.0984**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

Calinski-harabasz score (CH) [15], for evaluating the effectiveness of our method. Qualitative and quantitative experimental results are listed in Fig. 2 and Table 2, respectively.

In Fig. 2, it can be seen that all the five methods can find appropriate clusters due to a good distribution of data (i.e., high densities within a cluster and low boundary densities). Experimental results on data S1, S2, S3 and S4 illustrate more noise points using DPC with noise, regarding the maximum density of boundary points between each two clusters to be a threshold for noise. Due to large thresholds, inner points with smaller densities are viewed as noise. As to DBSCAN, noise is considered as the points which are not core points and cannot not be connected to core points. The parameter *MinPts* is used to measure core points. The bigger *MinPts* is, the less core points are obtained. As a result, more noise is illustrated. On the contrary, clusters are more likely to be wrongly merged due to a smaller *MinPts*. Our approach can divide points into more appropriate clusters, for noise can be recognized more properly using GMM-based model selection on densities other than a hard threshold. Besides, DPC, DPC-HD and Fuzzy-DPC achieve better performance for neglecting the existence of noise.

In Table 2, DPC, DPC-HD and Fuzzy-DPC also achieve better DBI, SC and CH scores. When noise is considered, A-DPC rather than DPC with noise and DBSCAN shows good performance, for it achieves better SC and CH scores. However, DBI scores of A-DPC are inferior. The more noise is discovered, the smaller distances within each cluster are obtained. That is why DPC with noise and DBSCAN own smaller DBI scores.

Secondly, we select six data (i.e., Aggregation [16], Compound [17], Flame [18], D31 [19], R15 [19] and Spiral [20])

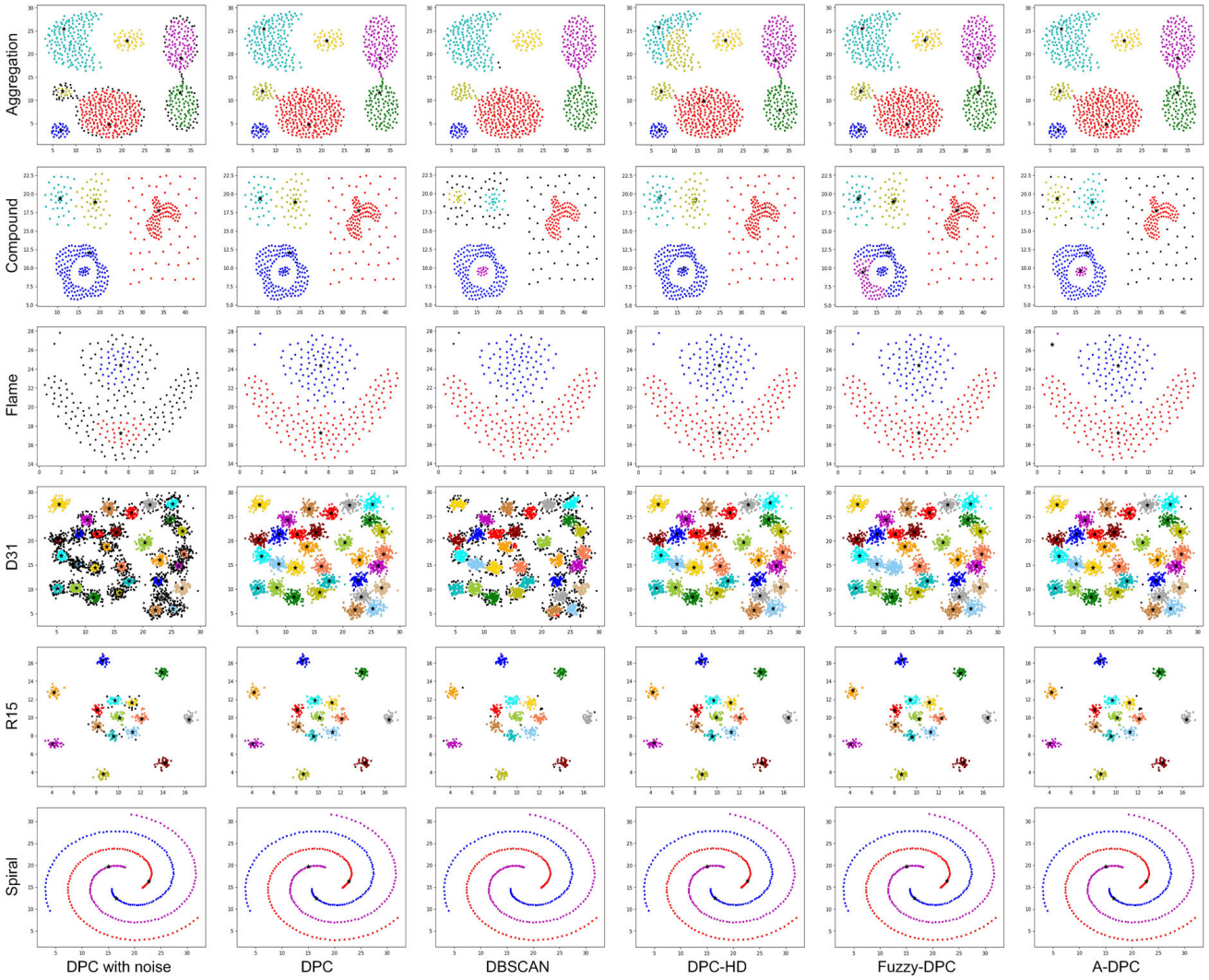


FIGURE 3. Qualitative results of clustering on six data with original labels.

for making qualitative and quantitative comparisons among DPC [1], DBSCAN [2], DPC-HD [6], Fuzzy-DPC [7] and A-DPC. Since all the six data keep original labels, we select three external criteria, i.e., Mutual Information (MI) [21], Adjusted Rand Score (ARI) [22] and Fowlckes-Mallows Index (FMI) [23], for evaluating the effectiveness of our method. Qualitative and quantitative experimental results are illustrated in Fig. 3 and Table 3, respectively.

In Fig. 3, it can be seen that out A-DPC almost obeys the original distributions for keeping the same number of clusters and eliminating the right noises; whereas, clustering by other methods doesn't. Immediate experimental results can be seen on Compound using DPC, DBSCAN, DPC-HD, Fuzzy-DPC and our A-DPC, respectively.

In Table 3, the clustering results on Flame, D31 and R15 using our A-DPC are a little different with the original sample distributions. Without considering noise, DPC performs the best experimental results. As to Flame shown in the third row of Fig. 3, two scatters on the top left corner of the space are labeled to the cluster with blue color, which makes A-DPC achieve slightly lower MI, ARI and FMI scores. Anyway, our approach still makes better quantitative clustering results with larger MI, ARI and FMI values in general.

Thirdly, we select two more data (i.e., Pathbased [20] and Jain [24]) for making qualitative and quantitative comparisons among DPC [1], DBSCAN [2], DPC-HD [6], Fuzzy-DPC [7] and our A-DPC. Since both of the two data continue to keep original labels, the three external criteria, i.e., Mutual Information (MI) [21], Adjusted Rand Score (ARI) [22] and Fowlckes-Mallows Index (FMI) [23], are still selected for evaluating the effectiveness of our method. Qualitative and quantitative experimental results are illustrated in Fig. 4 and Table 4, respectively.

Both qualitative and quantitative analyses illustrate poor results, due to small sample size which leads to an unbalanced

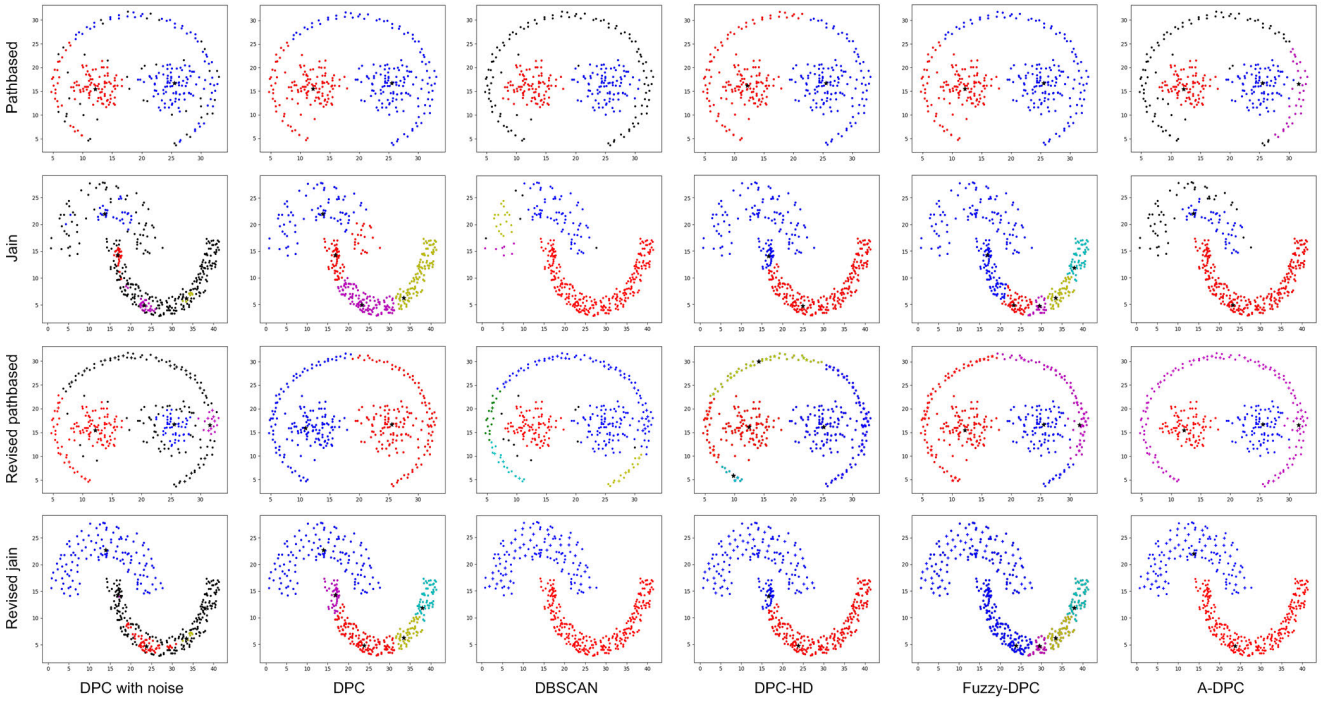


FIGURE 4. Qualitative results of clustering on two more data with original labels.

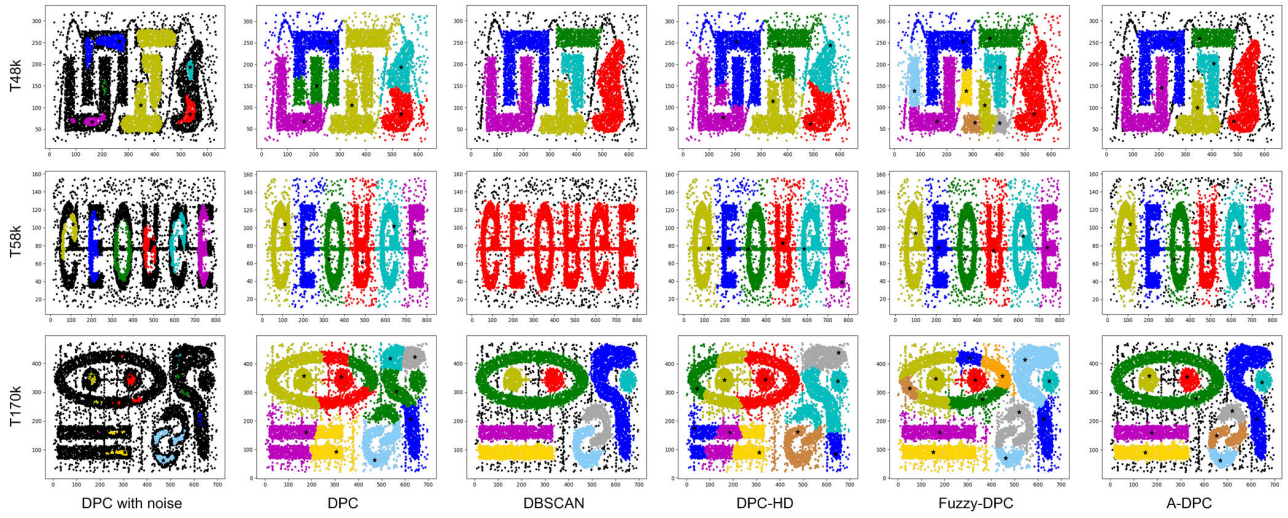


FIGURE 5. Qualitative results of clustering on three more data without original labels.

density distribution. In order to make density equalized, we supply 25 and 31 points (expressed as "+") to Pathbased and Jain, respectively. The corresponding qualitative and quantitative results are listed in the third and fourth row of Fig. 4 and Table 4. It can been seen that our method keeps the best results, which demonstrates the effectiveness of our method.

From Fig. 2 to Fig. 4, it can be concluded that A-DPC is an automatic density-based clustering method, which is robust to noise and sensitive to unbalanced density distributions among clusters. In order to further confirm these facts, additional experiments are performed on three more data [25] (i.e., T48k, T58k and T170k) for making further comparisons among DPC [1], DBSCAN [2], DPC-HD [6], Fuzzy-DPC [7] and A-DPC. Qualitative experimental results are illustrated in Fig. 5.

In Fig. 5, our A-DPC shows robustness to noise. Both A-DPC and DBSCAN achieve better results than DPC, DPC with noise, DPC-HD and Fuzzy-DPC on data T48k. As to data T58k, our A-DPC is the only method that not only

TABLE 4. Quantitative results of clustering on two more data with original labels.

Data	MI [21]			ARI [22]			FMI [23]		
Pathbased	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	0.4399	0.4566	0.8595	0.3834	0.4227	**0.8125**	0.6045	0.6688	**0.8758**
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	0.4389	0.4566	**0.9490**	0.3984	0.4227	0.7726	0.6517	0.6688	0.8449
Jain	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	0.4557	0.1564	**0.5731**	0.3025	0.0856	**0.9338**	0.6091	0.6400	**0.9737**
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	0.3649	0.2401	**0.5731**	0.6438	0.0017	0.9296	0.8502	0.4317	0.9720
Revised pathbased	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	0.4455	0.3275	0.6520	0.3226	0.3140	0.4834	0.5507	0.5561	0.6643
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	0.4846	0.4452	**0.8735**	0.4354	0.3846	**0.7461**	0.6375	0.6040	**0.8298**
Revised jain	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN	DPC + noise	DPC	DBSCAN
	0.6231	0.6231	**0.6307**	0.3953	0.6562	**1.0**	0.6560	0.8295	**1.0**
	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC	DPC-HD	Fuzzy-DPC	A-DPC
	0.4096	0.1647	**0.6307**	0.6723	-0.0496	**1.0**	0.8505	0.4970	**1.0**

FIGURE 6. Cluster analysis of Olivetti Face Database [26] using our method.

can successfully accomplish clustering but also can eliminate the effect of noise. Due to unbalanced density distributions, A-DPC almost achieves the appropriate clusters as DBSCAN does on data T170k.

From Fig. 3 to Fig. 5, it always can be discovered that unexpected clusters are generated using DPC-HD [6] and Fuzzy-DPC [7]. Although improvements have already been made on DPC [1], metric δ is still utilized. DPC-HD applies kernels to calculation of local densities in order to improve DPC. As to Fuzzy-DPC, cluster centers are automatically selected using hard thresholds of ρ and δ; moreover, temporary clusters are to be merged if one cluster resides at a d_c distance from other cluster with average density. In contrast, A-DPC works for discarding δ and automatically merging clusters considering density comparability between a boundary point and cluster centers.

Ultimately, we apply A-DPC to Olivetti Face Database [26] with its first 100 images under treatement. The similarity between two images is computed by following [27]. Qualitative experimental results are shown in Fig. 6. Here, faces with the same color belong to the same cluster, whereas gray images are not assigned to any cluster. Besides, cluster centers are labeled with white circles. It can been seen that we successfully categorize 56 images compared with 38 recognized images using DPC [1], which demonstrates the effectiveness of our method.

V. CONCLUSION AND PERSPECTIVE

In this paper, we make a new clustering method (namely A-DPC) for automatic find of density peaks. Focusing only on local densities, A-DPC is separated from DPC which also considers the metric which measures the minimum distance between a cluster center and a point with a higher density. All the samples are sequenced with their densities in descending order. Then, we categorize densities using GMM-based model selection considering the one-dimensional characteristics of densities, which helps to measure the density comparability between a boundary point and some cluster center(s) of its neighboring point(s). On assumption that boundary points which keep comparable densities with cluster centers should be regarded as inner points, an algorithm is proposed to traverse all the samples for automatically finding clusters and combining those with similarities simultaneously. Experimental results show that A-DPC is robust to noise and sensitive to unbalanced density distributions among clusters.

Apart from unbalanced density distributions among clusters, d_c is also an important condition for judging densities and connectivity. Different selection of d_c may affect cluster

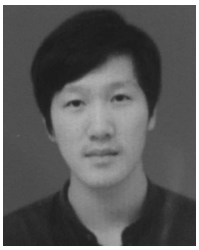
results. When d_c is small enough, all points will be isolated, which makes each point be a cluster. As d_c becomes large enough, all points will be grouped into one cluster. According to experience, we generally take the first 2% of all distances as DPC does. How to set this parameter more reasonably will be discussed in the near future.

ACKNOWLEDGMENT

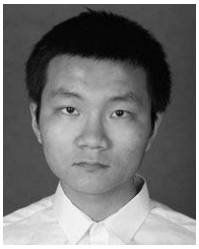
The funding body of Fundamental Research Funds for the Central Universities played an important role in the design of the study, collection, analysis, and interpretation of data and in writing the manuscript.

REFERENCES

[1] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014. doi: 10.1126/science.1242072.
[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, Oregon, Portland, 1996, pp. 226–231.
[3] X. Xu, M. Ester, H.-P. Kriegel, and J. Sander, "A distribution-based clustering algorithm for mining in large spatial databases," in *Proc. 14th Int. Conf. Data Eng.*, Orlando, FL, USA, Feb. 1998, pp. 324–331.
[4] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in *Proc. 4th Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 1998, pp. 58–65.
[5] Y. Zhu, K. M. Ting, and M. J. Carman, "Density-ratio based clustering for discovering clusters with varying densities," *Pattern Recognit.*, vol. 60, pp. 983–997, Dec. 2016. doi: 10.1016/j.patcog.2016.07.007.
[6] R. Mehmood, G. Zhang, R. Bie, H. Dawood, and H. Ahmad, "Clustering by fast search and find of density peaks via heat diffusion," *Neurocomputing*, vol. 208, pp. 210–217, Oct. 2016. doi: 10.1016/j.neucom.2016.01.102.
[7] R. Bie, R. Mehmood, S. Ruan, Y. Sun, and H. Dawood, "Adaptive fuzzy clustering by fast search and find of density peaks," *Pers. Ubiquitous Comput.* vol. 20, no. 5, pp. 785–793, Oct. 2016. doi: 10.1007/s00779-016-0954-4.
[8] Z. Li and Y. Tang, "Comparative density peaks clustering," *Expert Syst. Appl.*, vol. 95, pp. 236–247, Nov. 2017. doi: 10.1016/j.eswa.2017.11.020.
[9] R. Liu, H. Wang, and X. Yu, "Shared-nearest-neighbor-based clustering by fast search and find of density peaks," *Inf. Sci.*, vol. 450, pp. 200–226, Jun. 2018. doi: 10.1016/j.ins.2018.03.031.
[10] H. Q. Yan, L. Wang, and Y. G. Lu, "Identifying cluster centroids from decision graph automatically using a statistical outlier detection method," *Neurocomputing*, vol. 329, pp. 348–358, Nov. 2018. doi: 10.1016/j.neucom.2018.10.067.
[11] T. Huang, H. Peng, and K. Zhang, "Model selection for Gaussian mixture models," *Statistica Sinica*, vol. 27, no. 1, pp. 147–169, Jan. 2017. doi: 10.5705/ss.2014.105.
[12] P. Fränti and O. Virmajoki, "Iterative shrinking method for clustering problems," *Pattern Recognit.*, vol. 39, no. 5, pp. 761–775, May 2006. doi: 10.1016/j.patcog.2005.09.012.
[13] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979. doi: 10.1109/TPAMI.1979.4766909.
[14] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Nov. 1987. doi: 10.1016/0377-0427(87)90125-7.
[15] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist.*, vol. 3, no. 1, pp. 1–27, Sep. 1974. doi: 10.1080/03610927408827101.
[16] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," in *Proc. 21st Int. Conf. Data Eng.*, Tokyo, Japan, 2005, pp. 341–352.
[17] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," *IEEE Trans. Comput.*, vol. C-20, no. 1, pp. 68–86, Jan. 1971. doi: 10.1109/T-C.1971.223083.
[18] L. M. Fu and E. Medico, "FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data," *BMC Bioinf.*, vol. 8, no. 1, Jan. 2017, Art. no. 3. doi: 10.1186/1471-2105-8-3.
[19] C. J. Veenman, M. J. T. Reinders, and E. Backer, "A maximum variance cluster algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1273–1280, Sep. 2002. doi: 10.1109/tpami.2002.1033218.
[20] H. Chang and D.-Y. Yeung, "Robust path-based spectral clustering," *Pattern Recognit.*, vol. 41, no. 1, pp. 191–203, 2008. Jan. 2008. doi: 10.1016/j.patcog.2007.04.010.
[21] S. Thodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Burlington, MA, USA: Academic, 2009, p. 844.
[22] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, Jan. 1985. doi: 10.1007/BF01908075.
[23] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *J. Amer. Statist. Assoc.*, vol. 78, no. 383, pp. 553–569, 1983. doi: 10.1080/01621459.1983.10478008.
[24] A. K. Jain and M. H. C. Law, "Data clustering: A user's dilemma," in *Proc. 1st Int. Conf. Pattern Recognit. Mach. Intell.*, Kolkata, India, 2005, pp. 1–10.
[25] G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, Aug. 1999. doi: 10.1109/2.781637.
[26] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Sarasota, FL, USA, Dec. 1994, pp. 138–142.
[27] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2385–2401, Nov. 2009. doi: 10.1109/TIP.2009.2025923.



TONG LIU was born in 1998. He is currently pursuing the bachelor's degree with the College of Information and Computer Engineering, Northeast Forestry University, under the supervision of X. D. Zhao. His research interests include pattern recognition, medical image processing, and machine learning.



HANGYU LI was born in 1999. He is currently pursuing the bachelor's degree with the College of Information and Computer Engineering, Northeast Forestry University, under the supervision of X. D. Zhao. His research interests include pattern recognition and bioinformatics.

XUDONG ZHAO received the B.S. degree in intelligent instrument, the M.S. degree in computer science and technology, and the Ph.D. degree in artificial intelligence and information processing from the Harbin Institute of Technology, Harbin, China, in 2003, 2007, and 2013, respectively. He was a Postdoctoral Fellow of computer science and engineering with The Chinese University of Hong Kong, in 2014. He is currently an Assistant Professor with the College of Information and Computer Engineering, Northeast Forestry University, Harbin. His research interests include feature selection, clustering, discovery of signatures for prognosis in different cancers, differential expression analysis on expression profiles, and medical image processing.