An improved density peaks method for data clustering

Abdulrahman Lotfi
Department of Computer Engineering
University of Kurdistan
Sanandaj, Iran
a.lotfi@eng.uok.ac.ir

Seyed Amjad Seyedi
Department of Computer Engineering
University of Kurdistan
Sanandaj, Iran
amjadseyedi@eng.uok.ac.ir

Parham Moradi
Department of Computer Engineering
University of Kurdistan
Sanandaj, Iran
p.moradi@uok.ac.ir

Abstract- Clustering is a powerful approach for data analysis and its aim is to group objects based on their similarities. Density peaks clustering is a recently introduced clustering method with the advantages of doesn't need any predefined parameters and neither any iterative process. In this paper, a novel density peaks clustering method called IDPC is proposed. The proposed method consists of two major steps. In the first step, local density concept is used to identify cluster centers. In the second step, a novel label propagation method is proposed to form clusters. The proposed label propagation method also uses the local density concept in its process to propagate the cluster labels around the whole data points. The effectiveness of the proposed method has been assessed on a synthetic datasets and also on some real-world datasets. The obtained results show that the proposed method outperformed the other state-of-the art methods.

Keywords-Data clustering; density peaks; Rand index, local density function.

I. INTRODUCTION

Data clustering belongs to the category of unsupervised classification algorithms and is a powerful approach for data analysis. The main aim of data clustering is to identify groups of similar objects. Clustering has found many applications in different areas such as text mining[1], image analysis [2], pattern recognition[3], feature selection[4, 5], recommender systems[6] and customer segmentations [7]. Up to now, several data clustering methods have been proposed in the literature. The clustering approaches can be either classified into hierarchical or partitioning methods[8]. The goal of hierarchical clustering methods is to partition data points into successively fewer structures. While, divisive methods, starts with a cluster which contains all of data points and then in each iteration, a cluster is selected to split into two smaller ones. This process continued recursively until each data point is in its own singleton cluster. While, the goal of partition-based methods is to decompose data points into a set of groups where the number of the resulting clusters should be defined by the user. Partition-based methods can be divided into several categories such as: hard clustering, soft computing, density-based and model based methods[9].

In density-based methods a specific probability distribution is used to assign each point to each cluster [10]. In this types of clustering methods, it is assumed that the overall data distribution is modeled by a mixture of several distributions. Thus, the main goal of these methods is to identify clusters and their associated parameters of the probability distribution. Density-based methods are designed for identifying non-convex cluster shapes. The idea behind these methods is to continue growing a given cluster until the number of data points in its neighborhood exceeds a predefined threshold value [9]. These methods are easy to understand and they do not limit themselves to the shapes of clusters. Up to now several density-based clustering methods have been proposed in the literature including; DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN, SNOB , MCLUST and DBCLASD [11-14]. A comprehensive survey regarding the density-based methods which analysis advantages and disadvantages of these methods is presented in [15]. We refer the readers to the literature for more information about the density-based clustering methods [14, 16-18].

Recently, the authors of [19] proposed a density peaks clustering algorithm called DPC. DPC is a robust density-based clustering algorithm that is able to identify non-spherical clusters and does not require predefining the number of clusters. The idea behind their method is that cluster centers are characterized by a higher density compared to their neighbors. DPC used the global structure of data which results in missing many clusters. Besides, this method could not be able to perform well on high dimensional data. Up to now, several researches have been proposed in the literature to overcome the shortcomings of the DPC method. More recently, in [20] a method called DPC-KNN is proposed to improve the performance of the DPC. DPC-KNN employs the idea of k-nearest neighbors (k-NN) to consider the local structure of data.

In this paper, a novel density peaks clustering method is proposed. The proposed method consists of two main steps. In the first step, the local density concept is used to identify cluster centers. Using this concept, a new score function is proposed to rank each data point and then top k data points are identified as cluster centers. Moreover, in the second step, a novel label propagation method is proposed to form final clusters. This method also employs the local density concept in its process to identify the label of data points. The proposed label propagation method is completely differing from that of DPC-KNN[20]. It should be noted that in DPC-KNN, after identifying cluster centers, each data point is assigned to its nearest center, in order to form uniform cluster shapes. While in the proposed label propagation method, the label of each data point is propagated to its neighbors based on their local density values which leads to identify non-Spherical cluster shapes.

In Section 2, the proposed method is described. Section 3, presents the experimental results in synthetic and real datasets

78-1-5090-3586-1/16/$31.00 ©2016 IEEE

take from UCI repository. Finally, the last section concludes the intending work.

II. PROPOSED METHOD

In this section the proposed density peaks clustering method called IDPC is briefly described. The proposed method consists of two main steps including; (1) Identifying cluster centers and (2) Forming clusters. Cluster centers are identified among those data points that have higher density compared to their neighbors. To this end, in this method two different measures are used to recognize cluster centers. Besides, in the second step, a label propagation distance method is used to cluster data points. Additional details regarding these steps can be found in their corresponding sections. The pseudo-code of the proposed method is given in Algorithm 1.

Algorithm1: IDPC Pseudo-code

Inputs: The Samples $\mathbf{X} \in \mathbb{R}_{\mathbf{N}\times\mathbf{M}}$
The Parameter **P**

Outputs: The label vector of cluster index: $\mathbf{y} \in \mathbb{R}_{\mathbf{N}\times\mathbf{1}}$

Method:

Step 1- Identifying cluster centers
1: Calculate distance matrix using Eq. (2)
2: Calculate ρ for each point using Eq. (1)
3: Calculate δ for each point using Eq. (3)
4: Plot decision graph and select cluster centers
Step 2: Forming clusters
5: Assign labels of cluster centers to nearest neighbor points
6: Clustering based on neighborhood matrix and density
7: Assign each remaining point to the nearest cluster center
8: Return **y**

A. Identifying cluster centers

In this step to identify cluster centers, it is assumed that the cluster centers are surrounded by neighbors with a lower local density and also these centers should have a relatively larger distance to the points with higher density. To employing these assumptions two main measures are used. The first is local density of each point and other one is its distance from points with higher density. These measures correspond to the mentioned two assumptions with respect to the cluster centers. A simple definition for local density is mean distance of each data point with its neighbors. The local density ρ_i for data point x_i is defined as follows:

$$\rho_i = \exp\left(-\left(\frac{1}{k}\sum_{x_j \in N(x_i)} d(x_i, x_j)^2\right)\right) \qquad (1)$$

where $x_i = [x_{1i}, x_{2i}, \ldots, x_{mi}]$ is a data point vector with m attributes and n is the number of points. $d(x_i, x_j)$ shows the Euclidean distance between data points x_i and x_j and computed as follows:

$$d(x_i, x_j) = \| x_i - x_j \|^2, \; if \; x_j \in N(x_i) \qquad (2)$$

where $N(x_i)$ denote k nearest neighbors of data point x_i.

In Eq. (1), k is the number of neighbors and calculated based on a percentage p of the number of points ($k = p \times n$).
On the other hand, another measure to define cluster centers is defined as the minimum distance between a data point and any other points with higher density. This measure is denoted by δ_i and is defined as follows:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_i<\rho_j} \left(d(x_i, x_j)\right), \; if \;\; \exists j \;\; s.t. \;\; \rho_i < \rho_j \\ \max\limits_{j} \left(d(x_i, x_j)\right) \qquad , \quad otherwise \end{cases} \qquad (3)$$

For each data point x_i, both ρ_i and δ_i is obtained using Eqs. (1) & (3) respectively. Then, a decision graph for these data points is plotted. In this graph, the x-axis shows the local density value of each data point while y-axis depicts the corresponding delta value (i.e. δ) for each data point. According to this graph, those points which have high local density and high delta values are considered as cluster centers. These points are also called as peaks because they have higher densities than the other points. Finally, the following score function is used to rank the data points:

$$score(x_i) = \delta_i . \rho_i \qquad (4)$$

The data points are sorted based on their corresponding score values and then those of top c data points are selected as cluster centers.

B. Forming clusters

In this step a novel label propagation method is proposed to form clusters based on identified cluster centers. This method identifies the cluster of each data point based on the following four steps:

1- A distinct label is assigned to each cluster center.

2- Each cluster center propagates its label to its k nearest neighbors as follows:

$$Label(x_i) = \begin{cases} Label(peak_j), if \;\; x_i \in N(peak_j) \\ \varnothing \qquad , \quad otherwise \end{cases} \qquad (5)$$

3- For each data point x_i which doesn't have any label, if its local density value ρ_i is lower than ρ_j (i.e. x_j a neighbor of x_i) then x_i takes the label of x_j. If the local density of more than one neighbor is higher than ρ_i, then the label of x_i is identified using the voting method.

4- Finally if a data point x_i doesn't have any label, the label of nearest cluster center is assigned to x_i.

Figs. 1(a)-1(c), simply show how the proposed method in the second step forms final clusters. Fig. 1(a) shows two identified cluster centers. First, these cluster centers are assigned a distinct label. These labels are specified by yellow and blue colors. From Fig. 1(b) it is clear that the neighbors of these centers take the center labels. Following the steps 3 & 4, all the other points take a distinct label. Finally, Fig. 1(c) shows the final identified clusters. It can be seen from the Fig. 1(c) that the proposed method can group data points

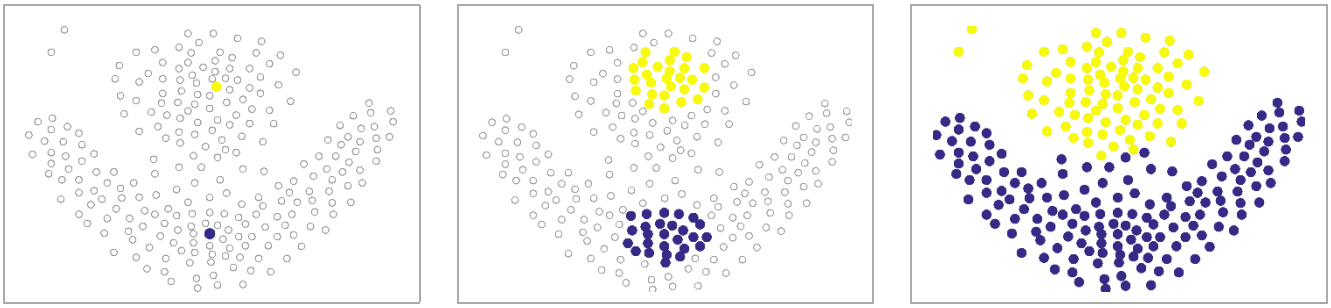

Fig. 1. (a) Assigning distinct label to peaks (b) assigning label of peak to its neighbors (c) propagating labels from labeled to unlabeled samples

into clusters with different shapes. It should be noted that the DPC-KNN[20] uses a different method to form clusters. In DPC-KNN after identifying cluster centers, the other data points take nearest cluster center associated label. While in the proposed method, a novel label propagation distance which uses the local density concept is used to identify final clusters.

III. Performance Evaluation

In this section we first demonstrate the feasibility of our algorithm on a synthetic dataset. Then the effectiveness of the proposed method is assessed on six real datasets taken from UCI repository. The additional details regarding these datasets are provided in *Table I*. The proposed method is also compared to well-known and state-of-the-art methods.

Table I. Details of real-world datasets taken from UCI.

Dataset	#instance	#feature	#cluster
Iris	150	4	3
Wine	178	13	3
Heart	270	13	2
Waveform	5000	21	3
Sonar	208	60	2
Glass	214	10	6

A. Evaluation metrics

In order to evaluate quality of the clustering results, in the experiments, two well-known evaluation metrics including; clustering accuracy (ACC)[21] and Rand index (RI) are used. The ACC metric is defined as follows:

$$ACC = \sum_{i=1}^{N} \delta(y_i, c_i) / N \qquad (6)$$

where N is the number of data points, $y_i$and c_i are the true label and the predicted label of data point x_i. Moreover, the Rand index is defined as follows:

$$RI = \frac{a+d}{a+b+c+d} \qquad (7)$$

where a and d are measures of consistent classifications while b and c are measures of inconsistent classifications [22].

B. Results on synthetic datasets

The proposed method is tested on four synthetic datasets which are generated to show three different geometric shapes. Fig. 2 shows the first dataset. This dataset contains 788 data points belong to 7 classes. The delta values(δ) and local density values (ρ) are calculated for each data point. Besides, Fig. 2(a) shows their corresponding decision graph. According to the decision graphs, cluster centers are those data points that have higher ρ and δ values compared to the others. From Fig. 2(b), it is clear that only seven data points meet the conditions and identified as cluster centers. This process is repeated for additional two datasets namely; R15 and D31 synthetic datasets. Each dataset and its corresponding decision graphs and also recognized cluster centers are shown in Figs. 2(c)-2(f). From the results it can be seen that the proposed method can identify cluster centers effectively.

C. Results of the experiments

Several experiments were performed to evaluate the performance of the proposed method (IDPC) with several well-known and state-of-the-art methods including, k-means and fuzzy c-means DPC-KNN and DPC-PCA methods on six real-world datasets of *Table I*. The obtained results are reported in *Tables II* & *III*. It can be seen from the results that in all cases except Heart dataset, IDPC achieved higher accuracy. The parameter p which was used in density based clustering methods (IDPC, DPC-KNN, and DPC-PCA) refers to a portion of data points, which is considered as the number of neighbors (i.e. k= p × n). For example the clustering accuracy of IDPC for the Waveform dataset was 0.794, while in this situation the clustering accuracy of DPC-PCA, DPC-KNN, FCM and k-means methods were 0.6452, 0.5840, 0.5119 and 0.5012.

Similar results were obtained based on RI metric and the results are reported in *Table III*. From the results it can be also seen that in most cases the proposed method outperformed the other aforementioned methods. The results of both ACC and RI metrics show that the proposed method obtained significant results over Waveform dataset compared to the other methods. Fig. 3. Plots the Waveform data and each class label is identified by a different color. Moreover, the decision graph of this dataset is shown in Fig. 4.

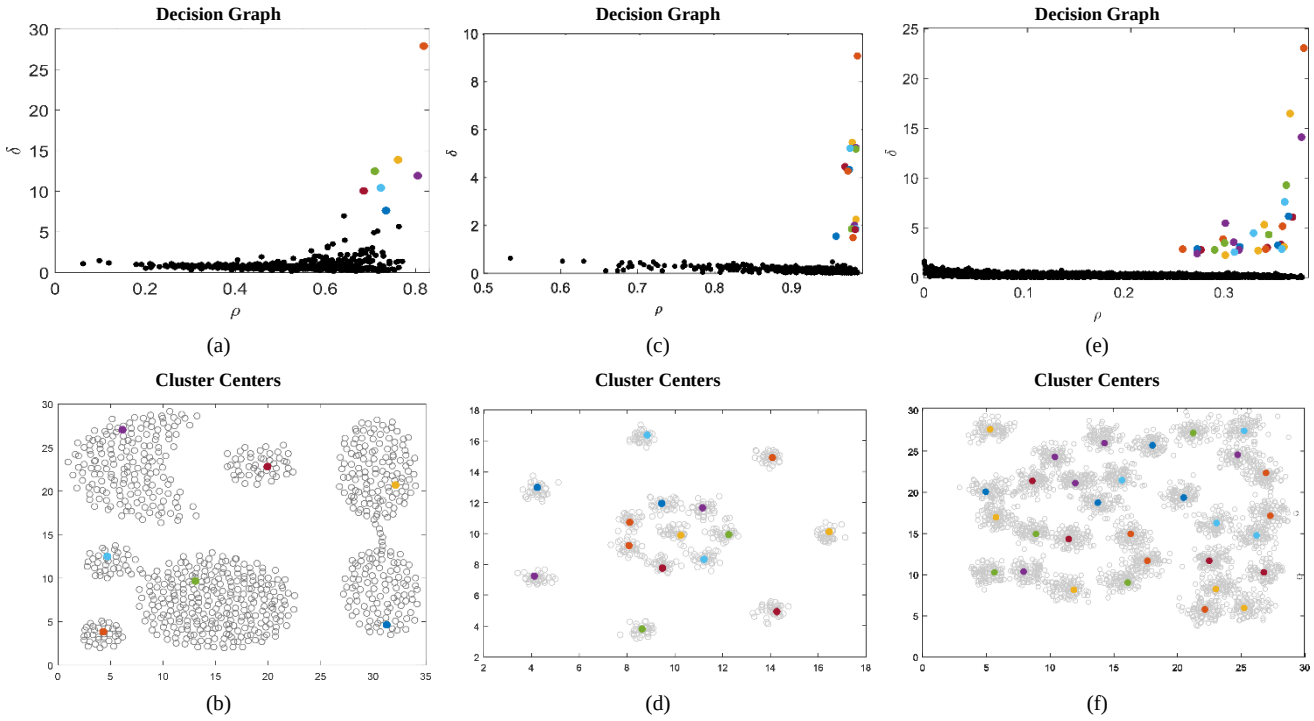

Fig. 2. (a) Decision graph of Aggregation dataset. (b) Aggregation cluster centers. (c) Decision graph of R15 dataset. (d) R15 cluster centers. (e) Decision graph of D31 dataset. (f) D31 cluster centers.

The results of Fig. 4 reveal that the decision graph distinguished three data points from the other points. Thus, these data points can be identified as cluster centers. On the other hand, from *Table I*, it can be seen that Waveform datasets consists of three classes. The obtained results of decision graph means that the proposed method could be able to identify true number of clusters for this dataset.

Table II. Comparison of ACC values obtained by different clustering methods.

Dataset	**Algorithms**				
	K-Means	**FCM**	**DPC-KNN**	**DPC-PCA**	**IDPC**
Iris	0.8560 ± 0.097	0.8733± 0.056	0.9266	0.9266	**0.94**
			p=8%	p = 9%	**P=21%**
Wine	0.5674±0 0.76	0.6853± 0.32	0.7247	0.7415	0.7415
			P=15%	P=0.6%	P=0.6%
Heart	0.7166± 0.0640	0.7625± 0.031	0.8111	**0.8259**	0.6925
			p=1%	**p=6%**	P=5%
Waveform	0.5012± 0	0.5119± 0.	0.5840	0.6452	**0. 794**
			p=0.2%	p=0.1%	**P=0.14%**
Sonar	0.5528± 0.22	0.571± 0.432	0.6105	0.6442	**0.649**
			P=4%	p=1%	**P=48%**
Glass	0.5514± 0.023	0.556± 0.055	0.5841	0.6962	**0.799**
			P=0.5%	P=0.5%	**P=3%**

Table III. Comparison of RI values obtained by different clustering methods.

dataset	**Algorithms**				
	K-Means	**FCM**	**DPC-KNN**	**DPC-PCA**	**IDPC**
Iris	0.8797± 0.023	0.8797± 0.055	0.9123	0.9123	**0.9495**
			P=9%	P=8%	**P=18%**
Wine	0.6854± 0.22	0.7105± 0.432	0.7360	0.7403	0.7403
			P=14%	P=0.6%	P=0.6%
Heart	0.5153±0. 172	0.520± 076	0.5642	**0.5669**	0.5358
			P=5%	**P=5%**	P=1%
Waveform	0.6672± 0	0.6599± 0	0.6720	0.7386	**0.7753**
			P=0.2%	P=0.1%	**P=0.14%**
Sonar	0.5032±0 76	0.5112± 0.432	0.5221	0.5066	**0.5243**
			P=4%	P=3%	**P=4%**
Glass	0.8486 ± 0.097	0.8545± 0.056	0.8146	0.8653	**0.8965**
			P=1%	P=30%	**P=38%**

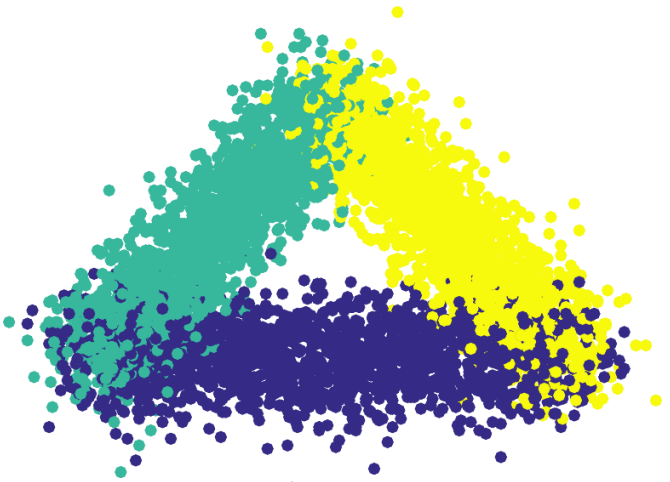

Figure 3. Waveform dataset with 3 classes

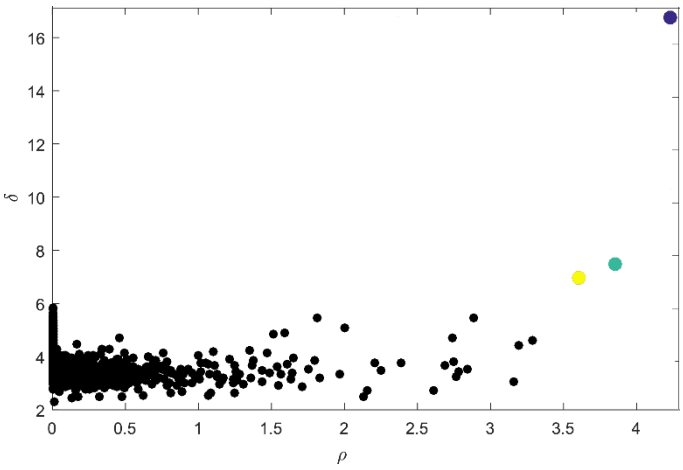

Figure 4. Decision graph of Waveform dataset

D. Sensitivity analysis of parameters

The proposed method contains only one parameter that needs to be properly set. This parameter is p which is used to define the number of neighbors in the clustering process. A small or large p leads to inappropriate cluster centers. Several experiments were performed on Waveform dataset to obtain a proper value for this parameter and the results are reported in Fig. 5. From the results it can be seen that IDPC obtained its highest accuracy when p is set to 0.0014.

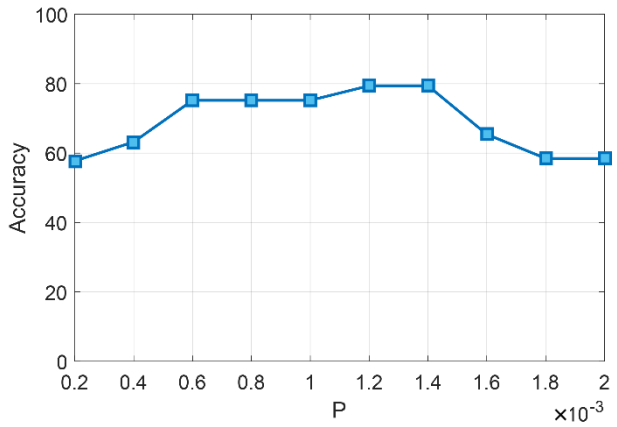

Figure 5. Performance of clustering on Waveform dataset by different P

IV. CONCLUSION

In this paper a novel density peaks clustering method was proposed. The proposed method called IDPC including two main steps. In the first step, the cluster centers are identified by using the local density concept. In the second step, a novel label propagation method was presented to i clusters. The proposed label propagation method, which is also based on local density concept is applied on each data point to find its proper cluster labels. This process is continued while all of data points are assigned by a cluster label. The performance of IDPC was evaluated on a synthetic datasets and also on some real-world datasets. The obtained results show that the proposed method obtained significant results compared to the other state-of-the art methods.

REFERENCES

[1] J. Wang and G. Karypis, "On efficiently summarizing categorical databases," Knowledge and Information Systems, vol. 9, pp. 19-37, 2006.

[2] M. Omran, et al., "Image classification using particle swarm optimization," in Proceedings of the 4th Asia-Pacific conference on simulated evolution and learning, 2002, pp. 18-22.

[3] U. Maulik and I. Saha, "Modified differential evolution based fuzzy clustering for pixel classification in remote sensing imagery," Pattern Recognition, vol. 42, pp. 2135-2149, 2009.

[4] P. Moradi and M. Rostami, "Integration of graph clustering with ant colony optimization for feature selection," Knowledge-Based Systems, vol. 84, pp. 144-161, 8// 2015.

[5] P. Moradi and M. Rostami, "A graph theoretic approach for unsupervised feature selection," Engineering Applications of Artificial Intelligence, vol. 44, pp. 33-45, 9// 2015.

[6] P. Moradi, et al., "An effective trust-based recommendation method using a novel graph clustering algorithm," Physica A: Statistical Mechanics and its Applications, vol. 436, pp. 462-481, 10/15/ 2015.

[7] D. S. Boone and M. Roehm, "Retail segmentation using artificial neural networks," International journal of research in marketing, vol. 19, pp. 287-301, 2002.

[8] S. Alam, et al., "An evolutionary particle swarm optimization algorithm for data clustering," in Swarm Intelligence Symposium, 2008. SIS 2008. IEEE, 2008, pp. 1-6.

[9] L. Rokach and O. Maimon, "Clustering Methods," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds., Boston, MA: Springer US, 2005, pp. 321-352.

[10] C. Fraley and A. E. Raftery, "Model-Based Clustering, Discriminant Analysis, and Density Estimation," Journal of the American Statistical Association, vol. 97, pp. 611-631, 2002/06/01 2002.

[11] C. S. Wallace and D. L. Dowe, "MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions," Statistics and Computing, vol. 10, pp. 73-83, 2000.

[12] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," The computer journal, vol. 41, pp. 578-588, 1998.

[13] M. Parimala, et al., "A survey on density based clustering algorithms for mining large spatial databases," International Journal of Advanced Science and Technology, vol. 31, pp. 59-66, 2011.

[14] M. Ester, et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," in Kdd, 1996, pp. 226-231.

[15] R. Bhuyan and S. Borah, "A Survey of Some Density Based Clustering Techniques."

[16] M. Ankerst, et al., "OPTICS: ordering points to identify the clustering structure," SIGMOD Rec., vol. 28, pp. 49-60, 1999.

[17] R. J. G. B. Campello, et al., "Density-Based Clustering Based on Hierarchical Density Estimates," in Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II, J. Pei, et al., Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160-172.
[18] J. Sander, et al., "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications," Data Mining and Knowledge Discovery, vol. 2, pp. 169-194, 1998.
[19] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," Science, vol. 344, pp. 1492-1496, 2014-06-27 00:00:00 2014.
[20] M. Du, et al., "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," Knowledge-Based Systems, vol. 99, pp. 135-145, 5/1/ 2016.
[21] E. Y. Chan, et al., "An optimization algorithm for clustering using weighted dissimilarity measures," Pattern Recognition, vol. 37, pp. 943-952, 5// 2004.
[22] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," Journal of the American Statistical Association, vol. 66, pp. 846-850, 1971.