Hindawi
Complexity
Volume 2021, Article ID 6666619, 12 pages
https://doi.org/10.1155/2021/6666619

Research Article

An Improved Integrated Clustering Learning Strategy Based on Three-Stage Affinity Propagation Algorithm with Density Peak Optimization Theory

Limin Wang,[1] Wenjing Sun,[2] Xuming Han ,[3] Zhiyuan Hao ,[4] Ruihong Zhou,[1] Jinglin Yu,[1] and Milan Parmar [2]

[1]*School of Internet Finance and Information Engineering, Guangdong University of Finance, Guangzhou 510520, China*
[2]*School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, Jilin, China*
[3]*College of Information Science and Technology, Jinan University, Guangzhou 510632, China*
[4]*School of Management, Jilin University, Changchun 130022, Jilin, China*

Correspondence should be addressed to Xuming Han; hanxvming@163.com and Zhiyuan Hao; 15391910163@163.com

Received 9 November 2020; Accepted 22 December 2020; Published 7 January 2021

Academic Editor: Abd E.I.-Baset Hassanien

To better reflect the precise clustering results of the data samples with different shapes and densities for affinity propagation clustering algorithm (AP), an improved integrated clustering learning strategy based on three-stage affinity propagation algorithm with density peak optimization theory (DPKT-AP) was proposed in this paper. DPKT-AP combined the ideology of integrated clustering with the AP algorithm, by introducing the density peak theory and k-means algorithm to carry on the three-stage clustering process. In the first stage, the clustering center point was selected by density peak clustering. Because the clustering center was surrounded by the nearest neighbor point with lower local density and had a relatively large distance from other points with higher density, it could help the k-means algorithm in the second stage avoiding the local optimal situation. In the second stage, the k-means algorithm was used to cluster the data samples to form several relatively small spherical subgroups, and each of subgroups had a local density maximum point, which is called the center point of the subgroup. In the third stage, DPKT-AP used the AP algorithm to merge and cluster the spherical subgroups. Experiments on UCI data sets and synthetic data sets showed that DPKT-AP improved the clustering performance and accuracy for the algorithm.

1. Introduction

Clustering analysis is an important research direction in the field of data mining. Through analyzing the internal structure information and spatial characteristics of massive data samples, the demand information can be obtained. Based on the advantages in data processing, clustering analysis is widely used in various fields of society. For example, clustering analysis can be used to identify viruses from a large number of virus research data; in artificial intelligence, it can be used for face recognition, fingerprint recognition, and other pattern recognition functions; in financial stock market, clustering analysis could be used to predict stock trends and so on. Therefore, how to improve the clustering method for meeting the research needs with massive data, obtain more accurate clustering results, and meet the demand-oriented user groups in the current society have become a hot issue and a key problem which need to be studied urgently by scholars all over the world [1–4].

In 2007, the American researchers put forward a novel clustering learning method named affinity propagation clustering algorithm in *Science*. The algorithm solves the problem of choosing the initial class representative point in the early stage of clustering. At the same time, there is no need to specify the clustering center, which largely avoids the risk that the algorithm will lead to local optimization due to the improper selection of parameters. However, the original AP algorithm still has some drawbacks as follows: it is unable to accurately deal with high-dimensional data, it needs to manually set the corresponding parameters, for some specific data types, it cannot accurately identify the data internal structure, and so on.

In this paper, in order to improve the clustering accuracy and clustering performance of the AP for the data with different structures and different sizes, in the AP algorithm, it introduced the density peak clustering theory and k-means algorithm and proposed the three-stage affinity propagation clustering algorithm based on combination optimization of density peak theory and k-means algorithm. The processes are listed as follows:

(1) This paper firstly used the density peak clustering algorithm to obtain the local density ρ and δ values and selected the values of $(\rho^* \delta)$ which are ranked in descending order, and the selection quantity of the $(\rho^* \delta)$ value is k.

(2) The k-means algorithm was used to carry on the secondary processing of data, through the DP algorithm determining the k clustering centers, and the data sample was divided into k subgroups

(3) Through AP clustering the subgroups, the new class label of the center point in a subgroup would be assigned to the other element point in the subgroup.

2. Related Works

With the arrival of the era of big data, the AP algorithm has become a very competitive clustering method in the field of data mining, and the applications of the AP algorithm are implemented in many different fields, for example, scholar E. Graham utilized the AP algorithm to propose a novel unsupervised clustering method in the microbial assemblies field [5]. Wang and Cheng introduced the affinity propagation to resolve the data-driven resource management issue for ultradense small cells [6]. Zhou and Xu combined the AP theory to resolve the issues of segmentation stability in the image segmentation field [7]. Aizpurua and Koutstaal utilized the affinity propagation clustering algorithm to research new index of semantic short-term memory and obtained better progress [8]. At the same time, scholar Chen et al. proposed a novel method for stability-based preference selection based on the AP algorithm [9]. Chinese scholars Zhang et al. extended the AP in a principled way to solve the image clustering problem and proposed the unsupervised image clustering method, which obtained the better result [10]. Ding et al. proposed a derived clustering algorithm for mixed-type data employing fuzzy neighborhood [11]. In the biology field, the scholar introduced the AP into the field of neuroscience data mining [12], etc. Also, in the other fields, a substantial number of scholars combined the affinity propagation clustering algorithm theory to handle the complex issues, including the tumor classification problem [13] and urinary-tract symptoms [14]. Because of the advantages of the AP, the application of the AP was accepted by numerous academics, and they introduced the theory of the AP into their research field to improve their original research results.

At the same time, in the original AP algorithm, there is a very important concept which is the similarity. And, it stipulates the Euclidean distance as the similarity calculation method for any two data samples. However, the Euclidean distance indicates the straight-line distance for any two points in a sample space. In view of the drawback of the Euclidean distance, when the AP algorithm analyzed the data set with intricate data framework, it cannot calculate the relevant precise similarity for the data points and finally obtain the inaccurate clustering result [15].

Given the similarity issue of the AP algorithm, many scholars proposed some different improvement algorithms. For example, Wang et al. altered the structure of the original algorithm to propose a novel self-adaptive affinity propagation clustering algorithm based on density peak theory and weighted similarity (DPWSAP). In the improved algorithm, it constructed a density attribution for the AP. Through weighting the density attribution and distance calculation method, the DPWSAP improved the similarity calculation accuracy, and finally, it obtained more accurate clustering results [16]. Wang et al. utilized the structure similarity to alter the original similarity calculation method to propose an adaptive semisupervised affinity propagation clustering algorithm (SAAP-SS). It started from the perspective of semisupervision, through the structure similarity, to handle a nonlinear, low-rank representation problem, then to improve the similarity calculation for data points, and finally to obtain the better clustering performance [17].

As it is known to all, there are two important parameters in the AP algorithm, including the *preference* and damping factor λ, and each parameter plays a momentous role in the clustering process. The *preferenc*e determines the final clustering numbers of the algorithm; when the value is selected higher, the final clustering number will be greater; also, when the value is selected smaller, the final clustering number will be fewer. For the clustering consequence, the suitable value of *preference* is more important. In view of this parameter, scholar Wang et al. proposed a density propagation-based adaptive multidensity clustering algorithm (DPAM), and the algorithm utilized a density propagation to reduce the impact of the parameter value and achieve the optimal clustering results [18]. Also, for the damping factor λ, the parameter can influence the convergence performance of the AP algorithm. In the clustering process, the suitable value of the damping factor can avoid the local optimal circumstance, in view of the different convergence speed of searching the clustering center in different stages; therefore, the value of the dynamic damping factor is very important. Considering the situation, Wang et al. combined the density peak algorithm and cut-off distance theory through these two theories to control the damping factor and improve the convergence performance of the original algorithm [19]. Wang et al. introduced the gravity concept to propose affinity propagation clustering algorithm based on gravity theory (GAP). GAP constructed a novel clustering method under the physical perspective. On the one hand, it improved the accuracy of similarity calculation for data points;

on the other hand, because of the improvement of algorithm structure, the GAP can control the convergence process of the algorithm well, and it reduced the impact of damping factor and improved the final clustering results [20].

From the above AP application and improved algorithms, we can learn that though the AP algorithm possesses the better application prospect, it owns some defects. The scholars introduced other research theories to improve the AP. However, these improvements have not changed the recognition performance on the data with different structures. They just made the AP obtain the relevant accurate clustering numbers. For the data samples with different shapes and densities, they could not obtain the better clustering result yet. Consequently, in order to improve the clustering accuracy and clustering performance of the algorithm for the data with different structures and different sizes, this paper introduced the DP algorithm and k-means algorithm and used three stages to cluster the data sample, and it improved the accuracy and efficiency compared with the original AP [21–24].

3. Theoretical Basis

3.1. Density Peak Clustering Algorithm. In 2014, the density peak clustering algorithm (DP) was proposed in *Science* , and compared with the early diversity clustering algorithms, the DP arose with a considerable breakthrough, and it greatly improved the performance of clustering algorithm [25]. In the literature [25], there are some merits in determining the final clustering centers. Through introducing the concept of cut-off distance and local density, the DP could apply to analyze the data samples with different types relatively better, including different densities and different shapes. And, the two parameters play the more important role in the clustering process. There are two assumptions making the DP effective [25]:

(1) The cluster center points are embraced by adjacent points with low local density

(2) For the data points with a larger local density, there is a relatively long distance between any two points

In the DP algorithm, a scientific cut-off distance d_c is adopted to calculate their local density ρ for sorting these density values in the descending order as follows [25]:

$$\begin{aligned} d_{ij} &= \operatorname{dist}\left(x_i, x_j\right), \\ \rho_i &= \sum_j \chi\left(d_{ij} - d_c\right), \\ \rho_i &= \sum_{j\neq i} \exp\left(-\left(\frac{d_{ij}}{d_c}\right)^2\right). \end{aligned} \tag{1}$$

When the algorithm creates the decision graph, there is a formula as follows:

$$\delta_i = \min_{j:\rho_j>\rho_i}\left(d_{ij}\right). \tag{2}$$

Formula (2) represents the minimum distance between the data point i and sample point with higher density. The decision graph is generated according to the ρ value and δ value which are obtained in the definition. As shown in Figure 1 [25], this is the distribution of data point with density size. And, in Figure 2 [25], the data point 10 and data point 1 have relatively high distance and local density at the same time, so they are clustering centers. However, the data points 26, 27, and 28 have relatively high distance, but the local density is smaller, so they are called outliers. For regular data points, the DP algorithm categorizes them into the category of the closest class center, that is, denser than theirs.

3.2. K-Means Algorithm. The k-means algorithm takes k as the parameter and divides n data objects into k classes. The data objects in each class have high similarity, but the similarity between different classes is relatively low. Similarity is calculated by calculating the average value of a data object in a cluster, and the definition of similarity is the key to division. The basic idea of the k-means algorithm is to randomly select k objects as the initial clustering center among n data objects; then, according to the principle of minimum distance, the distance from each data object to the clustering center is calculated and assigned to the nearest cluster. Then, the average value of each cluster is recalculated, and the convergence function is calculated until the center of each cluster no longer changes, and finally, the algorithm is terminated. Otherwise, the above process is repeated. The process of the k-means algorithm is shown in Table 1.

3.3. Affinity Propagation Clustering Algorithm. The core idea of the AP algorithm is to treat all sample points as potential class representative points and to minimize the decision function through the continuous transmission of two kinds of information: *availability* and *responsibility* so that the sample similarity within the cluster is the largest, and the sample similarity between different clusters is the smallest. Assume $\{x_1, x_2, \ldots, x_n\}$ to be a finite data set of the pattern space R_n, where x_i (i could have values of 1, 2, ... , ...) is a point composing of n-dimensional attributes, in a vector space. The similarity between any two samples s (i, k) is measured by a negative Euclidean distance [26] and is shown as follows:

$$S(i, k) = -\left\|x_i - x_k\right\|. \tag{3}$$

In the clustering process of the AP algorithm and before the two important information iterations, it needs to determine the value of parameter *preference*, which is s (k, k). This algorithm considers that the larger the value of the s (k, k), the more likely its corresponding point k is selected as the class representative point. In other words, the number of final clustering classes could be affected by the *preference* value. The affinity propagation clustering algorithm initially assumes that all data points could be chosen as potential class representative points with the same possibility, which is setting all s (k, k) to be the same *preference* value. Different *preference* values could result in different clustering results.

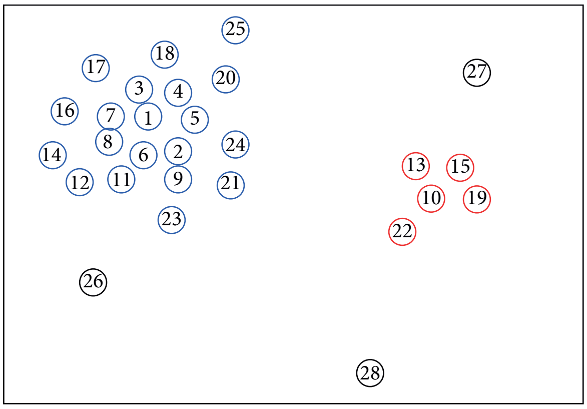


Figure 1: The distribution of the data point with density size.

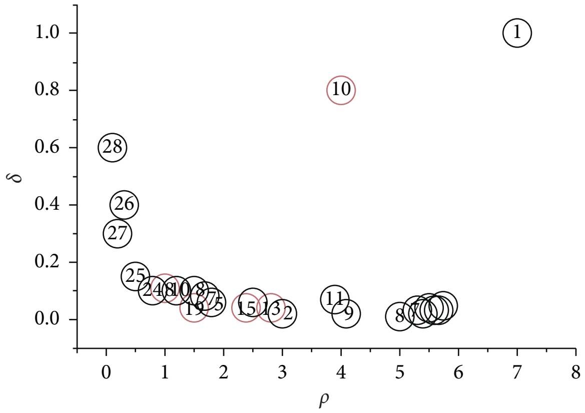


Figure 2: Information transfer between data points.

Generally, the AP algorithm selects the median or minimum value of similarity matrix to be the *preference* value [26].

The AP algorithm has two important information, which are the *responsibility* (r (i, k)) and *availability* (a (i, k)) mentioned above, and each kind of information is a competitive way of different representative points. They propagate continuously between any two data points and finally obtain a more reasonable clustering result. The *responsibility* and *availability* are constantly searched for in order to select suitable class representative points. For any sample point, in any iterative update stage, these two kinds of information together determine a certain sample point as a class representative point and which sample points belong to this class representative point. The iterative process of AP algorithm is actually the process of responsibility and availability alternatively updating the information. *Responsibility* indicates that the data point i sends the information to candidate class representative points k, reflecting the accumulated evidence of point k as cluster center of point i. At this time, there are many data samples competing with k point as the class center representative points of data point i. *Responsibility* is the information matrix which is established to select a final potential clustering center. *Availability* indicates that the candidate class representative point k sends the information to data point i, reflecting the accumulated evidence of the possibility for data point i selecting point k as its cluster center. Also, there are other points selecting the candidate class representative point k as their cluster center, and *availability* is also the information matrix which is established for this competitive mechanism [26].

At the beginning, assuming the value of a (i, k) equal to 0, two information updates are as follows:

$$a(i,k) \leftarrow \begin{cases} \min\left\{0, r(k,k) + \sum\limits_{i^{'} s.t.i^{'} \notin \{i,k\}} \left\{0, r\left(i^{'},k\right)\right\}\right\} i \neq k \\ \sum\limits_{i^{'} s.t.i^{'} \neq k} \left\{0, r\left(i^{'},k\right)\right\} \quad i = k \end{cases},$$
$$r(i,k) = s(i,k) - \max_{k' s.t.k' \neq j} \left\{a\left(i, k^{'}\right) + s(i,k')\right\}. \tag{4}$$

Through the updating iteration process of two kinds of message, *responsibility* and *availability,* between data sample points, the decision matrix E determines k as the final class representative point and is as follows:

$$E(k) = \arg\max_{k}(a(i,k) + r(i,k)). \tag{5}$$

The whole affinity propagation clustering algorithm can use the computer to calculate the two important similarities quickly and then obtain some reasonable numbers of clustering class. The above formulas determine any data sample point i could be the possible class center point in the case of the point i equal to point k. Also, the algorithm will eventually terminate because two kinds of information, *responsibility* and *availability,* are less than a certain threshold, or the local iteration situation does not change.

On the contrary, an important parameter, which is named as damping factor λ, is introduced in the updating of the affinity propagation clustering algorithm to avert numerical oscillation. During iteration, the renovating results of r (i, k) and a (i, k) can be obtained by computing the previous iteration results in each cycle iteration. Damping factor influences the convergence performance of the AP algorithm. When the number of classes generated by the AP algorithm continuously oscillates during iteration and cannot converge, increasing the damping factor can eliminate this oscillation. The range of damping factor values is [0, 1], and the default value is 0.5. The iteration process is as follows:

$$\begin{aligned} r^{(t+1)}(i,k) &\leftarrow (1-\lambda)r^{(t+1)}(i,k) + \lambda r^{(t)}(i,k), \\ a^{(t+1)}(i,k) &\leftarrow (1-\lambda)a^{(t+1)}(i,k) + \lambda a^{(t)}(i,k). \end{aligned} \tag{6}$$

4. Research Method

In order to improve the clustering accuracy and clustering performance of the algorithm for the data with different types and different sizes, this paper introduced the DP algorithm and K-means algorithm into the AP algorithm and

Table 1: The process of the k-means algorithm.

Step 1: randomly select k objects as the initial clustering center among n data objects

Step 2: according to the principle of minimum distance, the distance from each data object to the clustering center is calculated and assigned to the nearest cluster

Step 3: the average value of each cluster is recalculated, and the convergence function is calculated until the center of each cluster no longer changes

Step 4: when the cluster center does not change, the algorithm is over; otherwise, it will turn to Step 2

used three stages to cluster the data sample [27–32]. And, the stages are as follows:

(1) In the first stage, the clustering center point was selected by density peak clustering. Because the clustering center is surrounded by the nearest neighbor point with lower local density and has a relatively large distance from other version points with higher density, it could help the k-means algorithm in the second stage avoid the local optimal situation. The process is as follows:

The DP algorithm, firstly, coped with the data sample and obtained the local density ρ and δ values of each point. At the same time, the paper calculated the value of $(\rho^{*}\delta)$ and ranked the product value in descending order. According to the theory of the DP algorithm, the greater the product value is, the more likely the point is to become a class center. Thus, the paper selected the K-potential class centers by the product value from large to small.

(2) In the second stage, the k-means algorithm was used to cluster the data samples to form several relatively small spherical subgroups. Each subgroup has a local density maximum point, which is called the center point of the subgroup. The process is as follows:

Assume $P = \{p_1, p_2, p_3, \ldots, p_n\}$ is the data point set and $G = \{g_1, g_2, g_3, \ldots, g_k\}$ is the K subgroups which are obtained by the k-means algorithm. The value of K is from the first stage, and the center point of each subgroup is actually the potential clustering center point which is selected in the first stage. $D_K = \{\boldsymbol{D_1}, \boldsymbol{D_2}, \boldsymbol{D_3}, \ldots, D_i, \ldots, \boldsymbol{D_k}\}$ is the distance matrix, which indicates the distance between the elements in the subgroup and the K center points. There are K columns and n_k rows in the D_k. n_k is the number of elements of column j and also is the number of the elements in subgroup g_k. The paper made the distance of any two subgroups as follows:

$$\text{distance}\left(g_i, g_j\right) = \min\left(\min\left(D_i j\right), \min\left(D_j i\right)\right). \tag{7}$$

In formula (7), $D_i j$ is the all value of the column j in distance matrix D_i; $D_j i$ is the all value of the column i in distance matrix D_j, and n_j is the number of elements of column i. This paper used the distance which is defined in formula (7), rather than the distance between the any two center points. The calculation method is to find classes with nonconvex shapes, and formula (7) could provide more information about the compactness of two subgroups.

(3) In the third stage, because the AP algorithm is suitable for dealing with spherical data sets, based on this, the paper used the AP algorithm to merge and cluster the spherical subgroups formed in the second stage and finally realized the clustering analysis process of data samples. Experimental results show that the clustering accuracy of the DPKT-AP algorithm is obviously improved, and the clustering effect is better. The process is as follows.

The AP algorithm used the distance between any two subgroups as the similarity calculation method:

$$S(i, j) = \text{distance}\left(g_i, g_j\right). \tag{8}$$

The process of the DPKT-AP algorithm is in Table 2.

5. The Analysis of Simulation Experiment

To test the feasibility and effectiveness of the DPKT-AP algorithm, this paper compared it with the k-means, AP algorithm, and DP algorithm in three UCI data sets and two synthetic data sets listed in Table 3.

For proving the clustering accuracy of the developed DPKT-AP algorithm, this paper selected the three different algorithms which are the k-means, DP algorithm, and AP algorithm to compare with the DPKT-AP algorithm. According to the different densities and the different characteristics of the data sets to verify the clustering accuracy for the improved algorithm, we could use the clustering result to reflect the advantage of the DPKT-AP algorithm. The simulation experiment of the k-means, DP, original AP, and DPKT-AP algorithm was, respectively, tested in 5 different data sets. Comparing the four different clustering results, the following figures are clustering results. The paper could obviously obtain that through the three-stage clustering, and the DPKT-AP algorithm can obtain more accurate clustering numbers.

The subgroup center point of five different data sets is shown Figure 3, and as shown from Figures 4–8, the proposed DPKT-AP algorithm and the DP can aggregate clusters with varying structures and varying densities. The k-means and original AP algorithms cannot obtain the accurate clustering results. Flame and Aggregation belong to different structure data sets; Jain, D1, and D2 belong to different density data sets. For Flame and Aggregation data sets, the DPKT-AP and DP can detect classes of different shapes, and their results are almost the same. The original AP and k-means performed worse on Flame and Aggregation data sets. As for the original AP, no matter how it adjusts its parameters, it cannot find the correct clustering numbers on Aggregation data sets. More importantly, the results obtained by the AP are sensitive to the parameters *Preference* and Damping Factor, and the better results need to be carefully adjusted. For Jain, D1, and D2 data sets, they are made up of clusters of different shapes and densities. The DPKT-AP and DP found the correct clustering numbers on

Table 2: The process of the DPKT-AP algorithm.

Input: similarity matrix $S\ (i, j)$, cut-off distance d_c value, and initial parameter k
Output: final cluster number, division result $C = \{C_1, \ldots, C_k\}$, and the value of the evaluating indicators
Step 1: select d_c value
Step 2: density peak algorithm is used to calculate the local density ρ value and δ value
Step 3: according to the local density ρ value and δ value, the DP algorithm is used to get the initial clustering center point
Step 4: using the k-means algorithm to iterate the data sample and obtaining the several relatively small spherical subgroups, each subgroup has a local density maximum point, which is called the center point of the subgroup
Step 5: run the AP algorithm to go to the third stage of the clustering process, and use the evaluating indicators to evaluate the effectiveness of the algorithm

Table 3: The different data sets.

Data set	Sample number	Dimension	Class number
D1	**87**	**2**	**3**
D2	**85**	**2**	**4**
Jain	**373**	**2**	**2**
Flame	**240**	**2**	**2**
Aggregation	**788**	**2**	**7**

three data sets and almost obtained the same results. For D1 data set and D2 data set, the original k-means can get the correct clustering number; the AP could obtain the 3 classes and 5 classes, but they could not obtain the accurate sample data points' allocation. For Jain data set, the k-means algorithm could obtain the 2 classes, but the AP obtained the result with 3.

In this paper, in view of improving the clustering accuracy for the AP algorithm, it introduced the DP clustering and k-means algorithm into the original AP algorithm. The DPKT-AP combined the advantages of the DP, which is that the DP algorithm could find the center point quickly, and it has a relative advantage in identifying data with different sizes, densities, and shapes. And, the k-means could analyze the raw data to form spherical subgroups. From the above results, the proposed DPKT-AP algorithm obtains more improvements which are compared with the original AP algorithm. And, these improvements are mainly for the first two stages of the clustering process.

This paper used four different external evaluation methods to analyze the clustering performance of the compared algorithms, including Jaccard coefficient, Rand index, FM index, and $F1$ index. And, there are the following formulas of the four different evaluation methods [16, 20]:

$$M = a + b + c + d = \frac{N(N-1)}{2}. \tag{9}$$

In formula (9), a indicates the amount of data entity pairs which belong to the same class in the clustering results, but belong to different classes in the real structure; b indicates the amount of data entity pairs which belong to the same class in the clustering results and also belong to the same class in the real structure; c indicates the amount of data entity pairs which belong to different classes in the clustering results, but belong to the same class in the real structure; d indicates the amount of data entity pairs which belong to different classes in the clustering results and also belong to different classes in the real structure; N indicates the amount of all data entities [16, 20].

(1) Jaccard coefficient:

$$J = \frac{a}{a + b + c}. \tag{10}$$

(2) Rand index:

$$R = \frac{a + b}{M}. \tag{11}$$

(3) FM index:

$$\mathrm{FM} = \sqrt{\frac{a}{a + b}\frac{a}{a + c}}. \tag{12}$$

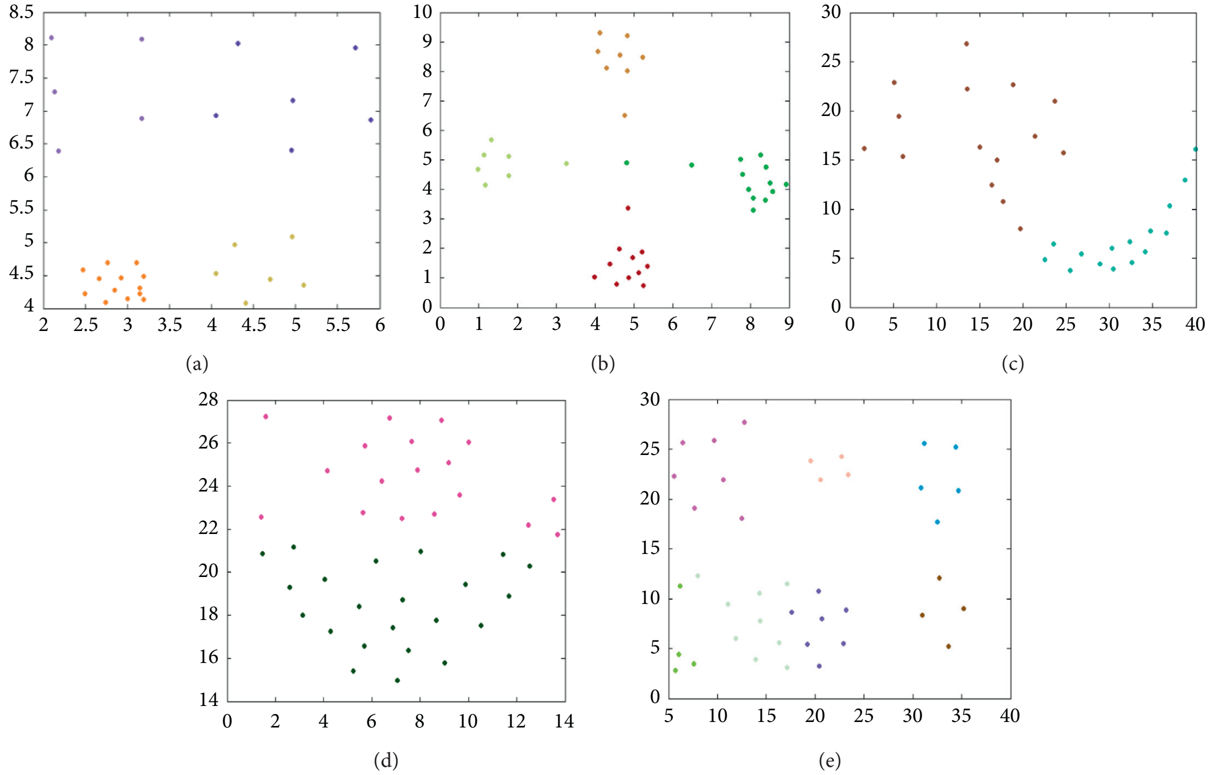
(4) $F1$ index:

$$\begin{aligned} P &= \mathrm{precision}(i,\ j) = \frac{N_{ij}}{N_i}, \\ R &= \mathrm{recall}(i,\ j) = \frac{N_{ij}}{N_j}. \end{aligned} \tag{13}$$

In formula (13), N_{ij} is the amount of classified i in cluster j; N_j is the amount of cluster j; N_i is the amount of classified i [16, 20]:

$$F1 = \frac{2PR}{P + R}. \tag{14}$$

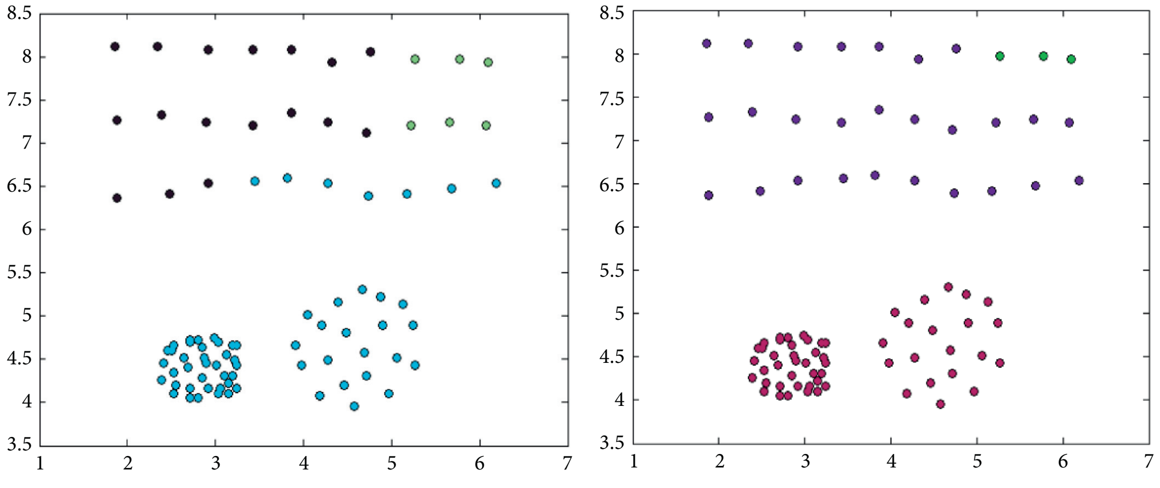
This paper utilized these evaluation indicator formulas to compare the AP, k-means, DP, and DPKT-AP algorithm. The result showed that the DPKT-AP algorithm is better among the four evaluation indicators. From Tables 4–7, there are evaluation results about the validity of the algorithm. The effectiveness of the algorithm evaluation results are listed in the following tables. From these four evaluation result tables, we can also get that the DPKT-AP algorithm can cluster data more accurately than the k-means algorithm and original AP algorithm, through the combination of the advantages between the k-means algorithm and DP algorithm. When the DPKT-AP processes data of different shapes and densities, it could obtain an apparent improvement of clustering performance, which is compared with the original AP, and it proves the theoretical feasibility for the DPKT-AP.



(a) (b) (c)

(d) (e)

Figure 3: The subgroup center point of five different data sets.



(a) (b)

Figure 4: Continued.

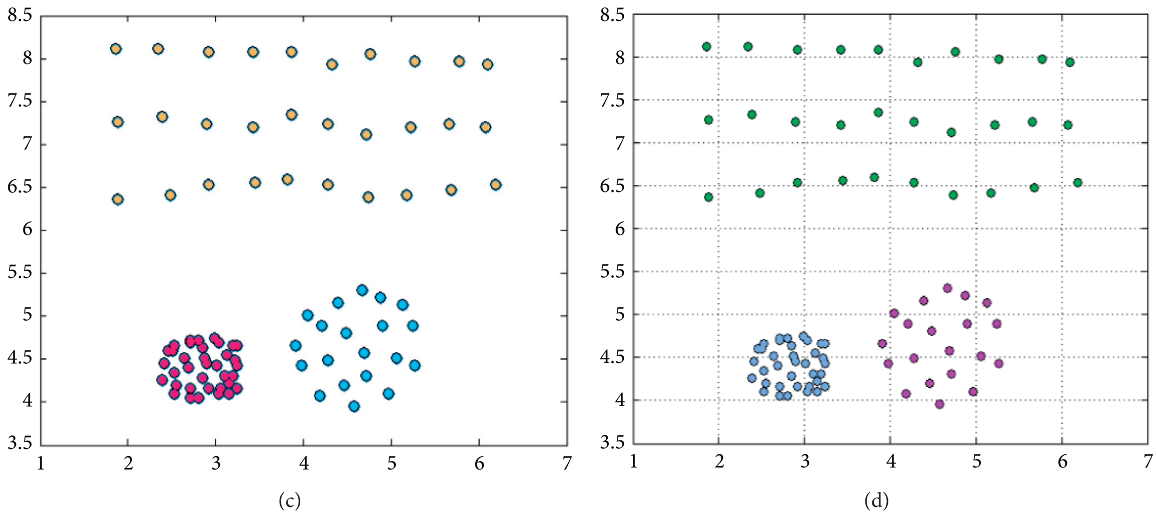


Figure 4: The clustering results of D1 data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.

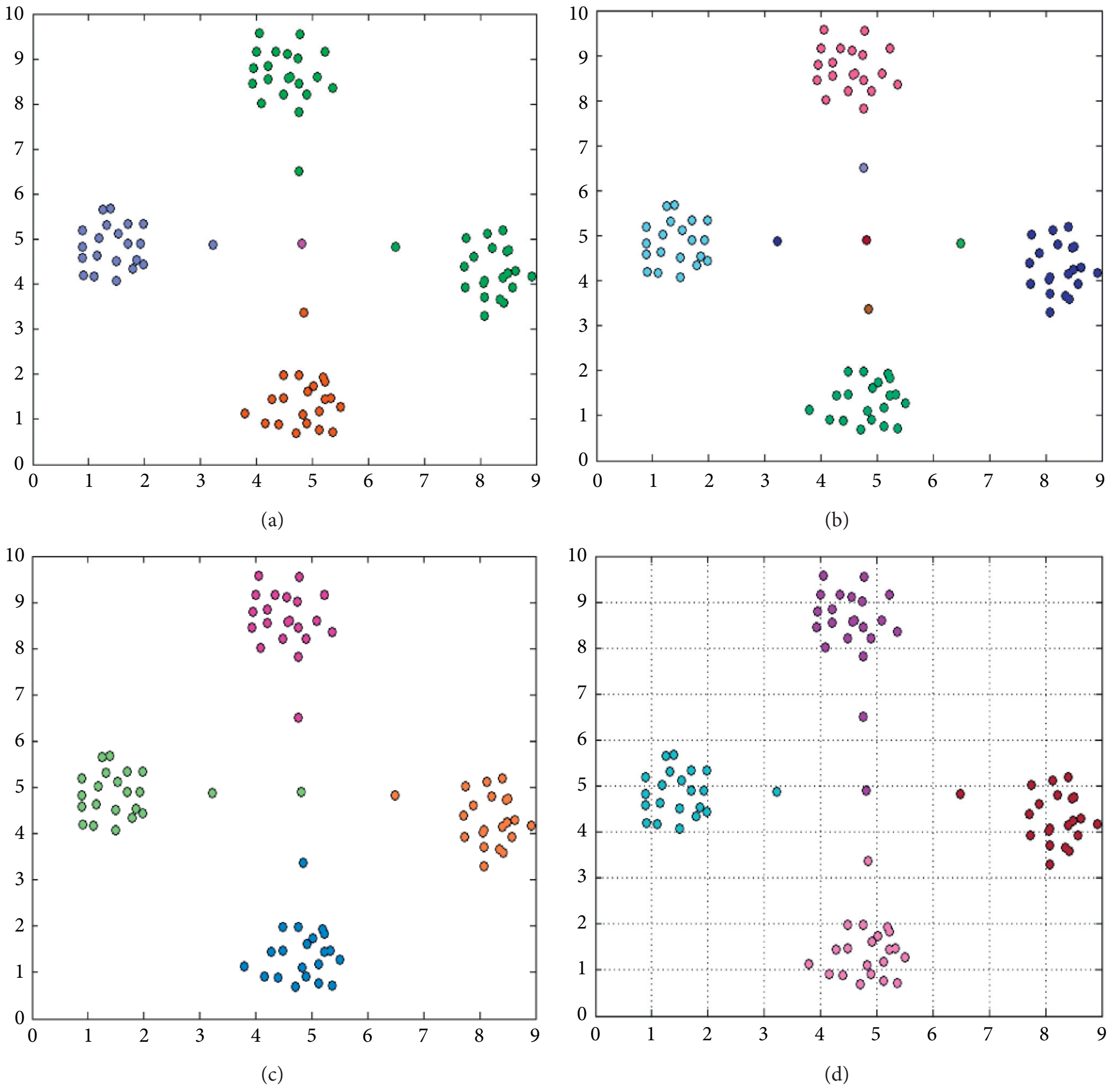
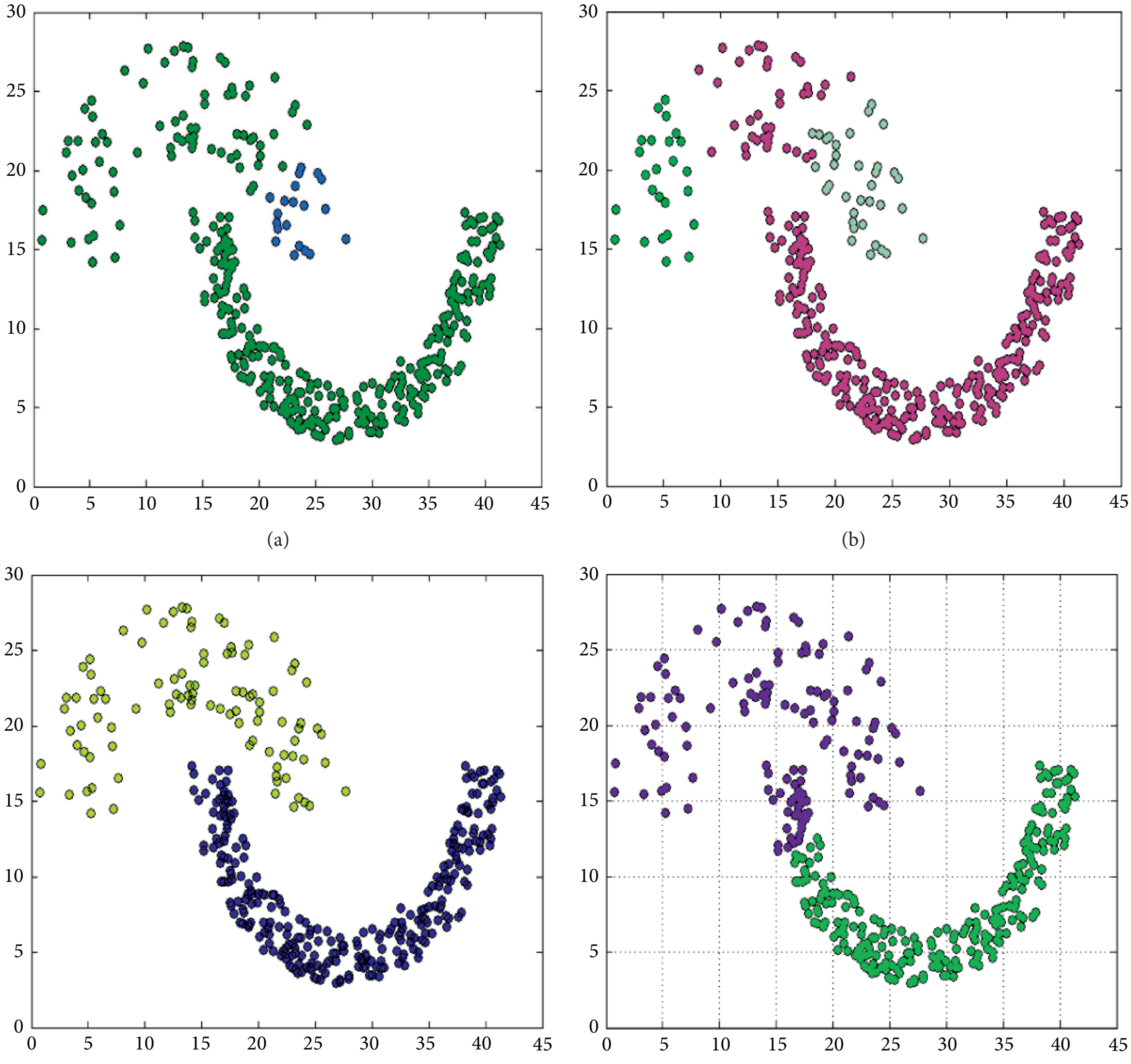
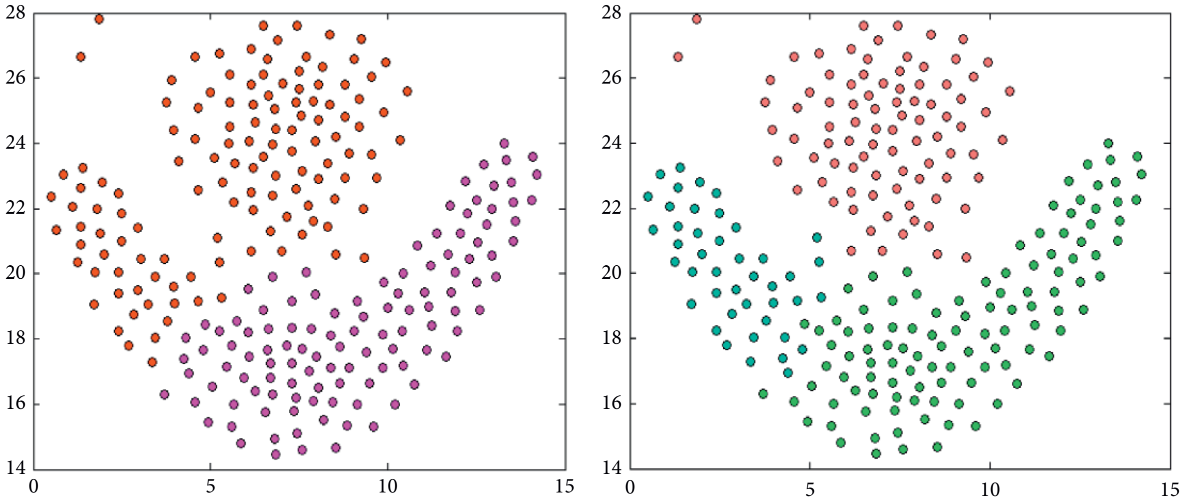


Figure 5: The clustering results of D2 data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.



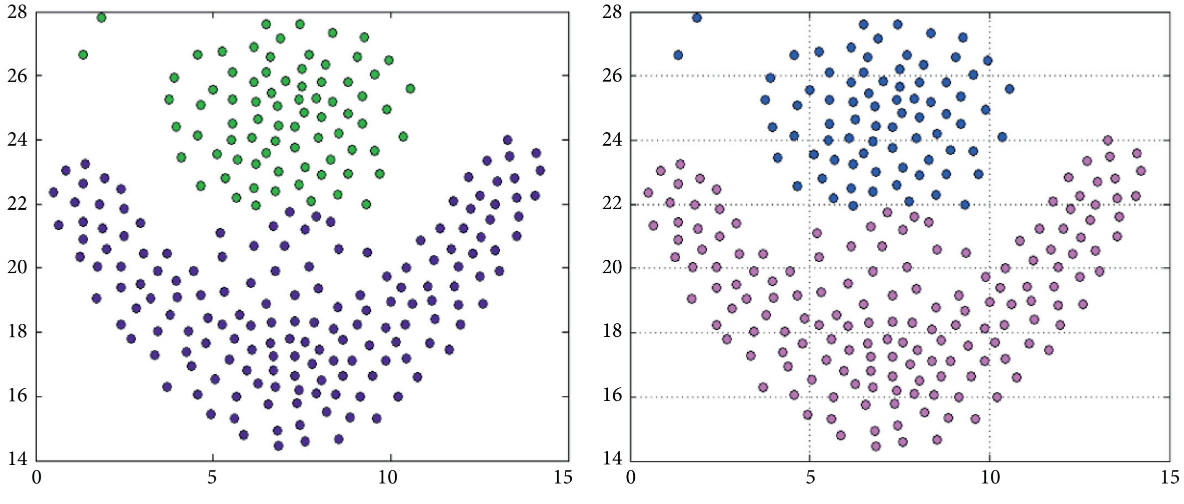
(a) (b) (c) (d)

Figure 6: The clustering results of Jain data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.



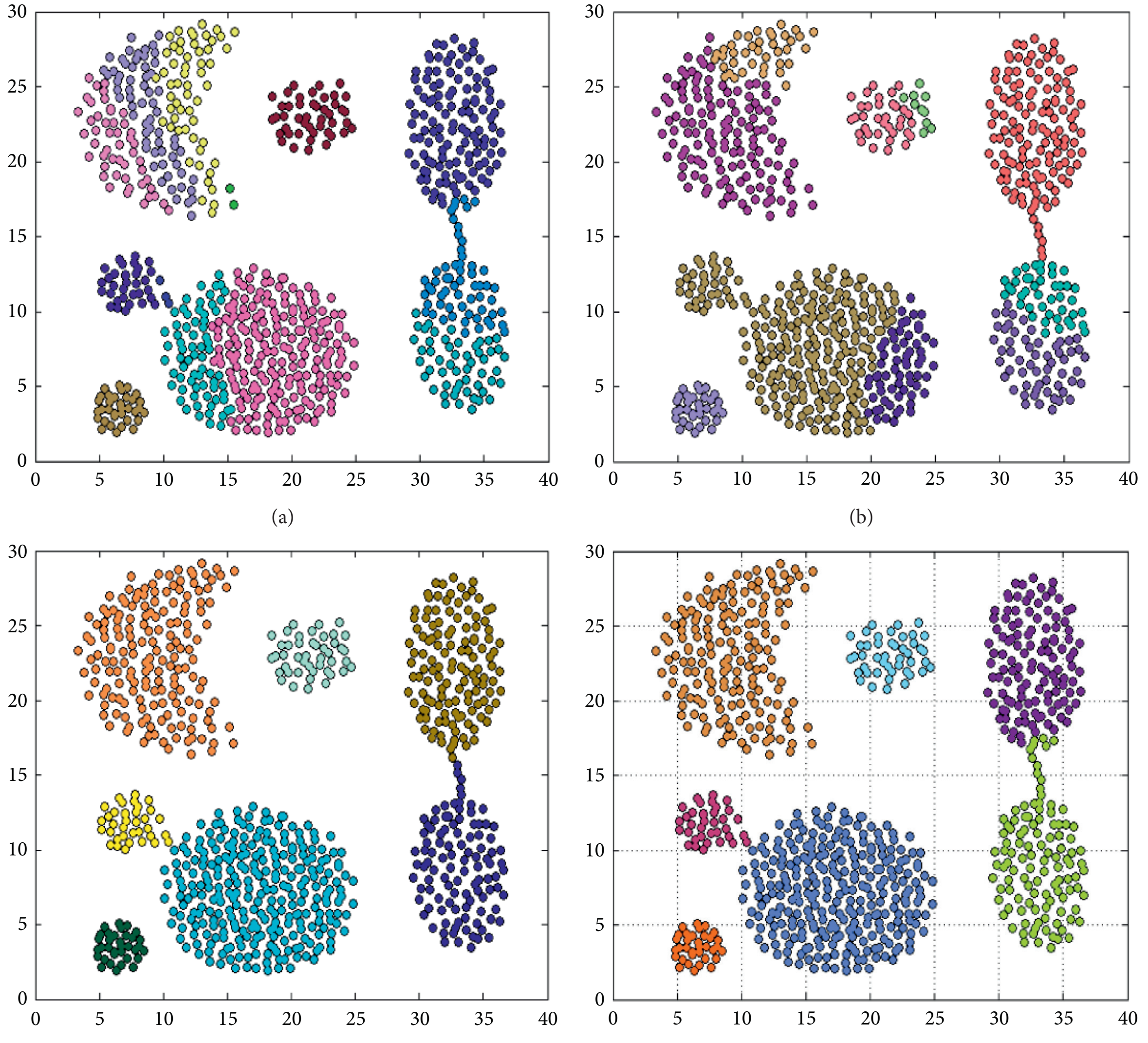
(a) (b)

Figure 7: Continued.



(c) (d)

Figure 7: The clustering results of Flame data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.



(a) (b)

(c) (d)

Figure 8: The clustering results of Aggregation data set in four different algorithms. (a) The result for the k-means. (b) The result for the AP. (c) The result for the DP. (d) The result for the DPKT-AP.

Table 4: The validity index value of the k-means.

Data set	K-means			
	FM	*F*1	Rand	Jaccard
D1	0.48	0.39	0.37	0.37
D2	0.45	0.41	0.33	0.28
Jain	0.57	0.49	0.41	0.39
Flame	0.29	0.27	0.19	0.19
Aggregation	0.35	0.33	0.24	0.21

Table 5: The validity index value of the DP.

Data set	DP			
	FM	*F*1	Rand	Jaccard
D1	0.81	0.79	0.71	0.68
D2	0.80	0.79	0.74	0.72
Jain	0.77	0.74	0.67	0.67
Flame	0.94	0.92	0.91	0.88
Aggregation	0.84	0.80	0.76	0.76

Table 6: The validity index value of the AP.

Data set	AP			
	FM	*F*1	Rand	Jaccard
D1	0.51	0.48	0.43	0.43
D2	0.49	0.48	0.41	0.39
Jain	0.55	0.53	0.47	0.41
Flame	0.54	0.50	0.44	0.83
Aggregation	0.61	0.69	0.81	0.88

Table 7: The validity index value of the DPKT-AP.

Data set	DPKT-AP			
	FM	*F*1	Rand	Jaccard
D1	0.79	0.76	0.76	0.71
D2	0.77	0.73	0.68	0.67
Jain	0.71	0.71	0.70	0.69
Flame	0.82	0.77	0.71	0.71
Aggregation	0.69	0.64	0.60	0.57

6. Conclusion

The outstanding contributions of this paper include combining the advantages of the DP algorithm and k-means algorithm with the original AP algorithm and proposing the improved integrated clustering learning strategy based on three-stage affinity propagation algorithm with density peak optimization theory (DPKT-AP). DPKT-AP has the advantage of high clustering accuracy. In view of that, the AP algorithm was suitable for processing spherical data, the DPKT-AP obtained the subgroups with spherical structures in advance by using the DP and k-means algorithms, and finally, the clustering process of the AP was carried out. Thus, better clustering results are obtained. Simulation results demonstrate that the DPKT-AP algorithm can reduce the difficulty of the clustering process for different size, structure, and density data sample and improve the clustering performance. Compared with the traditional algorithm, the proposed algorithm has obvious advantages. Of course, there are still some limitations in the proposed DPKT-AP, for example, higher time cost due to number of iterations and insufficient ability to identify outliers. In the future work, with regard to the situation that the clustering effect of high-dimensional data is weaker than the counterpart of lower-dimensional data and the remaining limitations, we will introduce a function which combines the density with distance or change the distance calculation method for the further study [32–37].

Data Availability

The data sets in the paper are available in http://cs.uef.fi/sipu/datasets/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Limin Wang and Wenjing Sun contributed equally to this work.

Acknowledgments

This study was supported by the National Science Foundation of China under Grant nos. 61572225 and 61472049, Foundation of Jilin Provincial Education Department under Grant no. JJKH20190724KJ, Jilin Province Science & Technology Department Foundation under Grant nos. 20190302071GX and 20200201164JC, and Development and Reform Commission Foundation of Jilin province under Grant no. 2019C053-11.

References

[1] V. Oona, "Using data mining methods to solve classification problems in financial-banking institutions," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 54, no. 1/2020, pp. 159–176, 2020.

[2] K.-S. Moon and H. Kim, "Performance of deep learning in prediction of stock market volatility," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 2, pp. 77–92, 2019.

[3] M. Miheala and C. Cristian, "Developing an index score for the internal auditor profile in Romania based on real data analysis," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 2, pp. 93–111, 2019.

[4] J. Yoon and S. Joung, "A big data based cosmetic recommendation algorithm," *Journal of System and Management Sciences*, vol. 10, no. 2, pp. 40–52, 2020.

[5] E. Graham, J. Heidelberg, and B. Tully, "BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation," *PeerJ*, vol. 5, 2017.

[6] L. Wang and S. Cheng, "Data-driven resource management for ultra-dense small cells: an affinity propagation clustering approach," *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 3, pp. 267–279, 2018.

[7] S. Zhou and Z. Xu, “Automatic grayscale image segmentation based on affinity propagation clustering,” *Pattern Analysis and Applications volume*, vol. 23, pp. 331–348, 2020.

[8] A. Aizpurua and W. Koutstaal, “A new index of semantic short-term memory: development and validation of the conceptual span task in Spanish,” *Plos One*, vol. 13, no. 12, 2018.

[9] D. Chen, J. Sheng, J. Chen, and C. Wang, “Stability-based preference selection in affinity propagation,” *Neural Computing and Applications*, vol. 25, no. 7-8, pp. 1809–1822, 2014.

[10] W. Zhang, X. Wu, W.-P. Zhu, and L. Yu, “Unsupervized image clustering with SIFT-based soft-matching affinity propagation,” *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 461–464, 2017.

[11] S. Ding, M. Du, T. Sun, X. Xu, and Y. Xue, “An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood,” *Knowledge-Based Systems*, vol. 133, pp. 294–313, 2017.

[12] H. K. Aljobouri, H. A. Jaber, O. M. Koçak, O. Algin, and I. Çankaya, “Clustering fMRI data with a robust unsupervised learning algorithm for neuroscience data mining,” *Journal of Neuroscience Methods*, vol. 299, pp. 45–54, 2018.

[13] L. Sun, X.-Y. Zhang, Y.-H. Qian, J.-C. Xu, S.-G. Zhang, and Y. Tian, “Joint neighborhood entropy-based gene selection method with Fisher score for tumor classification,” *Applied Intelligence*, vol. 49, no. 4, pp. 1245–1259, 2019.

[14] G. Liu, V. P. Andreev, M. E. Helmuth et al., “Symptom based clustering of men in the LURN observational cohort study,” *Journal of Urology*, vol. 202, no. 6, pp. 1230–1239, 2019.

[15] C. Yang, S. Liu, L. Bruzzone, R. Guan, and P. Du, “A feature-metric-based affinity propagation technique for feature selection in hyperspectral image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 5, pp. 1152–1156, 2013.

[16] L. Wang, Z. Hao, and W. Sun, “A novel self-adaptive affinity propagation clustering algorithm based on density peak theory and weighted similarity,” *IEEE Access*, vol. 7, pp. 175106–175115, 2019.

[17] L. Wang, Q. Ji, and X. Han, “Adaptive semi-supervised affinity propagation clustering algorithm based on structural similarity,” *Tehnicki Vjesnik-Technical Gazette*, vol. 23, no. 2, pp. 425–435, 2016.

[18] Y. Wang, W. Pang, Y. Zhou, R. Zhou, K. Zheng, and M. Liu, “Density propagation based adaptive multi-density clustering algorithm,” *Plos One*, vol. 13, no. 7, 2018.

[19] L. Wang, M. Li, X. Han, and R. Zhou, “Improved density peak clustering algorithm based on choosing strategy automatically for cut-off distance and cluster centre,” *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 2, pp. 536–545, 2018.

[20] L. Wang, Z. Hao, and X. Han, “Gravity theory-based affinity propagation clustering algorithm and its applications,” *Tehnicki Vjesnik-Technical Gazette*, vol. 25, no. 4, pp. 1125–1135, 2018.

[21] Z. Geng, R. Zeng, Y. Han, Y. Zhong, and H. Fu, “Energy efficiency evaluation and energy saving based on DEA integrated affinity propagation clustering: case study of complex petrochemical industries,” *Energy*, vol. 179, pp. 863–875, 2019.

[22] F. Xu, X. Shu, X. D. Zhang, and B. Fan, “Automatic diagnosis of microgrid networks’ power device faults based on stacked denoising autoencoders and adaptive affinity propagation clustering,” *Complexity*, vol. 2020, Article ID 8509142, , 2020.

[23] Z. Wei, Y. Wang, S. He, and J. Bao, “A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection,” *Knowledge-Based Systems*, vol. 116, pp. 1–12, 2017.

[24] L. Wang, X. Zhou, Y. Xing, M. Yang, and C. Zhang, “Clustering ECG heartbeat using improved semi-supervised affinity propagation,” *IET Software*, vol. 11, no. 5, pp. 207–213, 2017.

[25] A. Rodriguez and A. Laio, “Clustering by fast search and find of density peaks,” *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[26] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[27] Y. S. Park and J. H. Choi, “Algorithm of three-party combined judgment analysis engine for earthquake early warning system,” *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 45–64, 2019.

[28] J. B. Kim, “Implementation of artificial intelligence system and traditional system: a comparative study,” *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 135–146, 2019.

[29] K. Peng, V. C. M. Leung, and Q. Huang, “Clustering approach based on mini batch kmeans for intrusion detection system over big data,” *IEEE Access*, vol. 6, pp. 11897–11906, 2018.

[30] Y. Liu, Z. Ma, and F. Yu, “Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy,” *Knowledge-Based Systems*, vol. 133, pp. 208–220, 2017.

[31] B. Wu and B. M. Wilamowski, “A fast density and grid based clustering method for data with arbitrary shapes and noise,” *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1620–1628, 2017.

[32] L. Fu and Y. Dong, “Research on internet search data in China’s social problems under the background of big data,” *Journal of Logistics, Informatics and Service Science*, vol. 5, no. 2, pp. 55–67, 2018.

[33] C. Zhang, M. Ni, H. Yin, and K. Qiu, “Developed density peak clustering with support vector data description for access network intrusion detection,” *IEEE Access*, vol. 6, pp. 46356–46362, 2018.

[34] X. Xu, S. Ding, and Z. Shi, “An improved density peaks clustering algorithm with fast finding cluster centers,” *Knowledge-Based Systems*, vol. 158, pp. 65–74, 2018.

[35] M. Du, S. Ding, X. Xu, and Y. Xue, “Density peaks clustering using geodesic distances,” *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 8, pp. 1335–1349, 2018.

[36] A. Sajjad and F. Mehdi, “Particle swarm optimization algorithm for the prepack optimization problem,” *Economic Computation and Economic Cybernetics Studies and Research*, vol. 53, no. 2, pp. 289–307, 2018.

[37] J. Jiang, Y. Chen, D. Hao, and K. Li, “DPC-LG: density peaks clustering based on logistic distribution and gravitation,” *Physica A: Statistical Mechanics and Its Applications*, vol. 514, pp. 25–35, 2019.