IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON BIG DATA TECHNOLOGY AND APPLICATIONS IN INTELLIGENT TRANSPORTATION

Received November 14, 2019, accepted November 25, 2019, date of publication December 4, 2019,
date of current version December 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2956963

A Novel Self-Adaptive Affinity Propagation Clustering Algorithm Based on Density Peak Theory and Weighted Similarity

LIMIN WANG[1,2], ZHIYUAN HAO[3], AND WENJING SUN[1,2]

[1]School of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China
[2]Jilin Province Business Big Data Research Center, Jilin University of Finance and Economics, Changchun 130117, China
[3]School of Management, Jilin University, Changchun 130022, China

Corresponding author: Zhiyuan Hao (15391910163@163.com)

This work was supported in part by the National Science Foundation of China under Grant 61472049, Grant 61572225, and Grant 61202309, in part by the National Society Science Foundation of China under Grant 15bgl090 and Grant 19BJY246, in part by the Foundation of Jilin Provincial Science and Technology Research Project under Grant JJKH20190724KJ, and in part by the Society Science Research Foundation of Jilin Province under Grant 2019B69.

ABSTRACT To solve both the similarity calculation method and parameter limits problems of the affinity propagation algorithm (AP), the self-adaptive affinity propagation clustering algorithm based on density peak clustering and weighted similarity (DPWSAP) was proposed. The solutions were following: 1) density peak algorithm (DP) was introduced to create the local density attribute for AP algorithm; 2) weighted similarity was applied to heighten the similarity extent of data points; 3) growth curve function model was employed with setting a self-adaptive strategy for damping factor (λ) to enhance the convergence performance of AP at different stages. To verify the performance of DPWSAP we tested six UCI data sets with different density, different dimensions, and data volume. Experimental results indicated that DPWSAP had better clustering accuracy and convergence performance than original AP algorithm and several other clustering algorithms. In addition, the self-adaptive strategy improved the overall performance for the algorithm, and reduced the possibility of human factors affecting the algorithm effect. The analysis results demonstrated that the DPWSAP had a good research value. Thus, the proposed algorithm had a better research prospect in theory and application fields.

INDEX TERMS Affinity propagation, density peak theory, self-adaptive strategy, weighted similarity.

I. INTRODUCTION

Affinity propagation (AP) is a clustering algorithm based on data similarity calculation and it is proposed by American scholars in ***Science*** in 2007. The obvious advantage of AP is that the clustering center not need to be selected manually, the all data points could become potential clustering center, and also it can iterate constantly data samples with the process of running by the ***responsibility*** and the ***availability*** to get the optimal clustering center [1]–[9]. Being based on the characteristic, AP algorithm has been widely used in various kinds of data clustering analysis field. At the current, all over world scholars have made many improvements and continued a lot of research in the algorithm. For example, in the literature [10], Fujiwara and other writers in order to promote the convergence speed of affinity propagation algorithm, and on the premise of ensuring accuracy of clustering, they deleted the unnecessary information in the process of operation of the algorithm. In the literature [11], the scholars introduced the cuckoo search to improve the AP. Also in the literature [12], on the basis of the manifold learning thought, the writers had put forward an improved semi-supervised clustering algorithm, and the creation had improved the clustering performance of the algorithm. In the literature [13], the writers had put forward a self-adaption affinity propagation algorithm based on the singular value decomposition, the improvement can better solve the problem about the high dimensional data and further improve the clustering effect for the original affinity propagation algorithm.

Although the AP algorithm has many incomparable advantages over the other algorithms, with good and stable effect

The associate editor coordinating the review of this manuscript and approving it for publication was Sabah Mohammed.

in practical application, it still faces some difficulties and challenges:

(1) The AP algorithm does not need to specify the number of clusters and can get optimal cluster center gradually by information exchange, but the algorithm exist the parameter ***P*** *that* is set by the artificial selection to affect the final clustering result.

(2) The algorithm needs to obtain the final clustering center through continuous iteration, which leads to the high time complexity of the algorithm. And at the same time, the algorithm needs to set the damping factor to control the convergence of clustering results, which greatly increases the influenced of human factors on the final clustering results and limits the clustering performance for the algorithm.

(3) Since affinity propagation clustering algorithm mainly uses negative ***Euclidean*** distance between exemplars and samples as the similarity of them, it is difficult to identify clusters with complex structure, and the final clustering result was worse. Consequently, the performance of AP deteriorates on samples distributed with complex structure.

In view of this, on the premise of retaining the advantages of the traditional AP algorithm, it is of great significance to improve its shortcomings reasonably and effectively, improve the clustering performance of the algorithm, so that it can be more widely used in practical work and provide more effective decision-making basis for the government and enterprises.

The main contributions of this study are as follows:

(1) Considered the limitations of the traditional similarity calculation method, this paper combined DP algorithm, defined the local density of every point and constructed the density attribute, then weighted the distance and density for every data point to update the similarity matrix.

(2) At the same time, this paper defined a new function model based on the growth curve function to propose an adaptive damping strategy to improve the convergence performance of the algorithm in the global search and local search at different stages and obtained the best clustering results.

(3) Used the proposed method to compare with the AP algorithm and K-means algorithm in the clustering effect, and proved the proposed method was better than traditional clustering methods.

(4) The results of different algorithms under 6 different data sets were fully compared and analyzed. And the final experimental results showed that the convergence performance and clustering effect of the DPWSAP algorithm were obviously improved over the AP algorithm and K-means algorithm, and the application value of DPWSAP algorithm was good. Finally, the last section gave the conclusion and future research directions.

II. AFFINITY PROPAGATION ALGORITHM

Affinity propagation clustering algorithm is different from K-means clustering algorithm, K-means clustering algorithm is impacted on selecting the initial cluster point, it needs to repeatedly set different clustering initialization parameters in order to achieve a high quality of clustering results. On the other hand, the affinity propagation clustering algorithm treats all samples as potential cluster center, the samples continuously passed through the two kinds of information which are ***responsibility*** and ***availability***, each sample point will eventually find a group of their own class. The input value of affinity propagation is the similarity relation matrix ***S*** that is constructed by using the similarity of any two data points. The similarity is calculated by the ***Euclidean*** distance of two points [14], [15].

At the beginning of affinity propagation clustering algorithm, taking as input a real number ***s(k, k)*** for each data point ***k***, these values are named ***preferences***. These data points with larger values of ***s(k, k)*** are more likely to be chosen as exemplars. The number of clusters is influenced by the values of the input ***preferences***, the value of the input ***preference*** is even greater, the possibility of representative points is greater, and the number of clustering output is more numerous. If not, the value of the input ***preference*** will be smaller, and the number of clustering output will be less. If a prior, all data points are equally suitable as exemplars, so namely all the ***s(k, k)*** is the same value ***p***. In traditional affinity propagation clustering algorithm, the ***preference*** value is defined the median of the input similarities or their minimum.

In order to select the appropriate representative point, there are two kinds of messages exchanged between data points, namely ***responsibility*** and ***availability***, which each represents a different competitive goal. The ***responsibility r(i, k)***, means $\boldsymbol{x}_i$ point to candidate exemplar $\boldsymbol{x}_k$ that reflects the accumulated evidence for how suitable $\boldsymbol{x}_k$ is to serve as the exemplar for $\boldsymbol{x}_i$, taking into account other potential exemplars for $\boldsymbol{x}_i$. The ***availability a(i, k)***, means candidate exemplar $\boldsymbol{x}_k$ point to $\boldsymbol{x}_i$ that reflects the accumulated evidence for how appropriate it would be for $\boldsymbol{x}_i$ to choose $\boldsymbol{x}_k$ as its exemplar, taking into account the support from other points that $\boldsymbol{x}_k$ should be an exemplar. The larger ***r(i, k)*** and ***a(i, k)*** are, the larger the possibility that $\boldsymbol{x}_k$ is final class representative point. Affinity propagation is the iterative process that ***responsibility*** and ***availability*** update alternately [16]–[27].

With updating the messages, it is more significant that introducing the important parameter damping factor λ to avoid numerical oscillations that arise in some iterative circumstances. And the following formulas are the specific description for the process.

In affinity propagation algorithm, for any two points in the sample space between $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$, the similarity expressed in ***s(i, k)***. The mathematical expression is the following:

$$s(i,\ k) = -||x_i - x_k|| \tag{1}$$

In a priori, all data points are taken as the potential cluster centers. A data point with a large value of ***s(k, k)*** is more likely chosen as exemplar. These values are referred to as ***preference*** parameters. They play important roles in determining the number of exemplars.

$$p = median(s(:)). \tag{2}$$

The core of AP is the mutual transfer of the two information. The ***responsibility r(i,k)*** from point i to point k. It reflects how well-suited point k is to serve as the exemplar for point i. The ***availability a(i,k)*** from point k to point i. It reflects how appropriate it would be for point i to choose point k as its exemplar. From the view of evidence, larger the value of $r(:,k)$ and $a(:,k)$, more probability the point k as a final cluster center. A decision matrix E is calculated after each update. Decision matrix E represents whether point i chooses point k as its exemplary or not.

$$r(i,k) = s(i,k) - \max_{k's.t.k' \neq j} \left\{ a\left(i,\ k'\right) + s\left(i,\ k'\right) \right\} \tag{3}$$

$$a(i,k) \leftarrow \begin{cases} \min\left\{0, r(k,k) + \sum_{i's.t.i' \notin \{i,k\}} \max\left\{0, r(i',k)\right\}\right\} & i \neq k \\ \sum_{i's.t.i' \neq k} \max\left\{0, r(i',k)\right\} & i = k \end{cases} \tag{4}$$

$$r^{(t+1)}(i,k) \leftarrow (1-\lambda) r^{(t+1)}(i,k) + \lambda r^{(t)}(i,k), \tag{5}$$

$$a^{(t+1)}(i,k) \leftarrow (1-\lambda) a^{(t+1)}(i,k) + \lambda a^{(t)}(i,k), \tag{6}$$

$$E(k) = \arg\max_{k} \left(a(i,k) + r(i,k)\right). \tag{7}$$

III. ALGORITHM DESCRIPTION

A. THE CONSTRUCTION OF DENSITY ATTRIBUTION FOR ORIGINAL AP ALGORITHM

In the original AP algorithm, the input value is the similarity matrix, but the similarity is only calculated with the ***Euclidean*** distance between any two data points. The similarity based on ***Euclidean*** distance can not express the potential structural relationship accurately, and at the same time, the ***preference*** parameter was difficult to determine the accurate clustering numbers. The phenomenon could caused the final clustering results to be extremely unreasonable. Considerations based on these aspects, in this paper we introduced the density peak clustering algorithm (DP) to define a local density for all the points. ***Rho(i)*** was defined the density for the ***i*** point, and ***rho(j)*** was defined the density for the ***j*** point. In consideration of the ***responsibility r(i, j)*** and the ***availability a(i, j),*** we could obtain the core theory of the proposed algorithm from the following figures. At the first, the two data points sent messages to each other, and in the original AP algorithm, the similarity of the data point was calculated by the ***Euclidean*** distance of the point ***i*** and point ***j***,

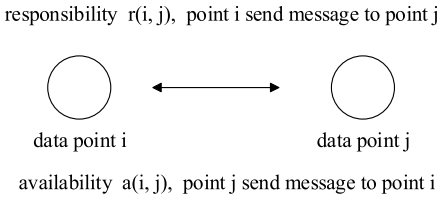


FIGURE 1. Any two data points send a message to go on the information transmission.

but we were not sure the point ***i*** can be the clustering center on the point ***j***, therefore, we should think about the following figure 2.

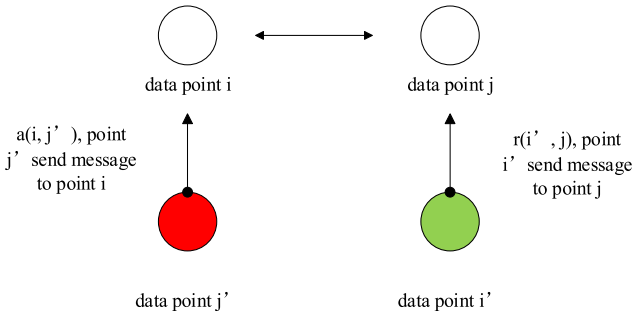


FIGURE 2. The other data points send a message to go on the information transmission.

The ***responsibility*** is the information that node ***i*** send to the node ***j***, conveys the attraction of node ***j*** to the node ***i***, which is recorded as ***r(i, j)***, then how to measure the attraction. In fact, the attraction is a relative concept, and we have a similarity matrix that records the degree to which ***j*** becomes the cluster center of ***i***, so in here, we only needed to prove that ***j*** is more appropriate than the other nodes. Then how did other nodes fit this measure, whether it was appropriate to see whether these two nodes agree with each other, and for other nodes j', we had $s(i,j')$ to indicate the suitability of node j' as the cluster center of node ***i***. And as we all know:

$$s(i,j) = r(i,j) + \max\left\{a\left(i,\ j'\right) + s\left(i,\ j'\right)\right\} \tag{8}$$

If the data point i' and j' affect the selection of the clustering center, thus the method using ***Euclidean*** distance to calculate the similarity is not accurate, based on this, we should consider the impact of information transfer on other data points. In this paper, we used the theory of density peak clustering algorithm to defined a local density for each data point, and took full account of the factors that affected the similarity, the process is Fig. 3.

From Fig. 3, we defined a density calculation method, and at the moment, we used the density and the ***Euclidean*** distance to calculate the similarity together. For any data point ***i*** and point ***j***, we could define a ***rho(i)*** and ***rho(j)***, also we could get the distance between ***i*** and ***j***. For the data point ***i***, we got the following formula (9), the ***availability*** (***a(i, j)***) in AP algorithm shows the suitability of j' to select point ***i*** as its

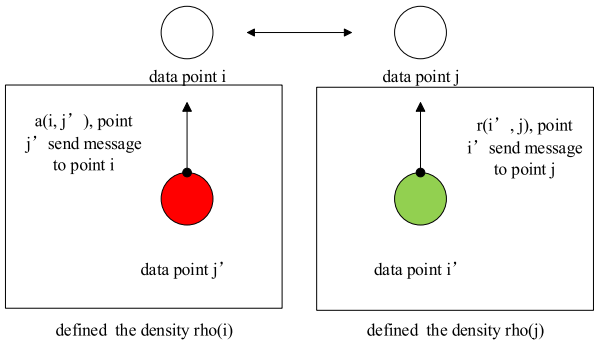


FIGURE 3. The definition of the data point density for AP algorithm.

clustering center, if the distance of any two points is enough small, we can believe the similarity is biggest.

$$\lim_{j' \to i} a\left(i,\ j'\right) = a\left(i,\ i\right) \to max\left(s\left(i, j'\right)\right) \tag{9}$$

This formal mean that when the data point $\boldsymbol{j'}$ is infinitely closed to the point $\boldsymbol{i}$, we can construe the influence of other points is closed to the minimum. In this paper, we introduced the ***cut-off*** distance of the DP algorithm, and in the DP algorithm, the local density depended on the value of ***cut-off*** distance parameter, and the density of the data point $\boldsymbol{i}$ can be calculated by the number of points in a range, when ***cut-off*** distance is smaller, the points in the range are more closed to the point $\boldsymbol{i}$, and also :

$$\lim_{j' \to i} s\left(i,\ j'\right) = s\left(i,\ i\right) \tag{10}$$

The number of neighbor points per data point is one percent to two percent of the total number. Then, we considered the second attribute that affects the similarity Now, we can define the next formula:

$$\lim_{j \to j} D\left(i, j'\right) = \alpha \times r\left(i, j'\right) + \beta \times \max\left\{a\left(i,\ j'\right) + s\left(i,\ j'\right)\right\} \tag{11}$$

In this formula, we set the two parameters to show the degree of influence for two attributes in similarity calculation.

$$s\left(i, j\right) = \alpha \times D\left(i,\ j\right) + \beta \times rho\left(i\right) \times rho\left(j\right) \tag{12}$$

The $\boldsymbol{D(i, j)}$ is the distance of the point $\boldsymbol{i}$ and point $\boldsymbol{j}$. Because the influence degree of two kinds of attributes on similarity is not able to get the specific proportion distribution, in this paper, therefore, we weighted the values of two attributes. Hence, we got the formula is the following:

$$\lim_{i' \to i, j' \to j} s\left(i,\ j\right) = sum\left(D\left(i,\ j'\right) .^{*} rho\left(i\right) .^{*} rho\left(j\right)\right) \tag{13}$$

$$Sum\left(D\left(i, :\right) < d_c\right) - 1 \tag{14}$$

But the ***rho(i)*** and ***rho(j)*** are based on that the distance is smaller than ***cut-off*** distance, therefore we could define the final similarity is the following:

$$S = \frac{Sum\left(\zeta^{*} D\left(i, j\right) .^{*} \left(\xi^{*} rho\left(i\right)\right) .^{*} \left(\varphi^{*} rho\left(j\right)\right)\right)}{Sum\left(D\left(i, :\right) < d_c\right) - 1} \tag{15}$$

$$\xi + \zeta + \varphi = 1 \tag{16}$$

According to the similarity calculation formulas, then the process of DPWSAP algorithm is shown in Table 1:

TABLE 1. Process of DPWSAP algorithm.

Input: Similarity matrix $S(i, j)$, Cut-off distance d_c value
Output: Cluster number $\boldsymbol{k}$, Division result C=$\{C_1, \cdots, C_k\}$ the value of the evaluating indicators
Step1: Select $\boldsymbol{d_c}$ value
Step2: Density peak algorithm is used to calculate density $\boldsymbol{\rho}$
Step3: Using the weighted similarity to obtain similarity matrix
Step4: Using the adaptive damping factor to update responsibility and availability
Step5: Using the final matrix to guide the clustering of AP algorithm and to obtain the clustering results
Step6: Run the AP algorithm, and use the evaluating indicators to evaluate the effectiveness of the algorithm
Step7: Record the clustering results.

B. CONSTRUCTION OF ADAPTIVE STRATEGY

The AP algorithm had two important parameters, one was the ***preference***, the number of clustering was greatly influenced by the ***preference***, and the second was the damping factor. The damping factor did not only affect the number of clustering but also played a decisive role in the convergence speed of the algorithm. And the improper selection of damping factors could lead to the oscillation of the algorithm, then it was impossible to converge, and finally it influenced the clustering effect. The damping factor of original AP algorithm was acquiescent a fixed value, and the value never changed during the operation of the algorithm. However, if the damping factor could change in the operation process of algorithm, the final clustering result could be more accurate. In every stage of the algorithm, the necessary for the damping factor was different, so the adaptive of the damping factor was more important. In view of this, this paper based on the growth curve function defined a new function model. We introduced a function method, proposed adaptive damping strategy, according to the convergence speed of each stage in the algorithm, used the function adaptively to adjust the damping coefficient, and improved the convergence performance of the algorithm in the global search and local search at different stages, finally obtained the best clustering results.

And in the original AP algorithm, the ***responsibility*** and ***availability*** would update, and every iteration was influenced

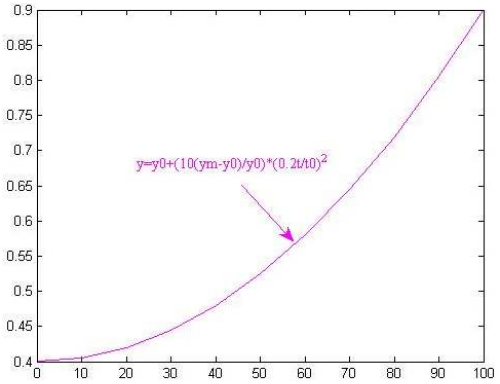


FIGURE 4. The search function image of damping factor in the previous period (formula 19).

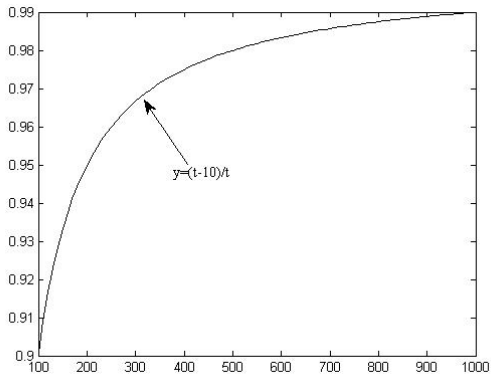


FIGURE 5. The search function image of damping factor in the later stage (formula 20).

by the damping factor, they were shown with the following formula.

$$r^{(t+1)}(i,k) \leftarrow (1-\lambda)r^{(t+1)}(i,k) + \lambda r^{(t)}(i,k), \quad (17)$$

$$a^{(t+1)}(i,k) \leftarrow (1-\lambda)a^{(t+1)}(i,k) + \lambda a^{(t)}(i,k), \quad (18)$$

Also in AP algorithm, the damping factor value acquiesced in 0.4. In the adaptive damping strategy, the initial damping coefficient set λ_0, the final value was λ_m, the maximum iteration number was $\boldsymbol{t}_m$, the current iteration number was $\boldsymbol{t}$, the iteration number of the function strategy change point was $\boldsymbol{t}_0$, then the adaptive dynamic damping strategy function is shown as follows.

$$\lambda(t) = \lambda_0 + \frac{10(\lambda_m - \lambda_0)}{\lambda_0} * (\frac{0.2^* t}{t_0})^2 \quad 0 \le t \le 100 \quad (19)$$

$$\lambda(t) = \frac{t-10}{t} \quad 100 \le t \le 1000 \quad (20)$$

And the following were the function figure. The first figure was the iteration number from 0 to 100, and the second figure was iteration number from 100 to 1000.

This paper set the initial value $\boldsymbol{\lambda_0}$ to be 0.4, the final value $\boldsymbol{\lambda}_m$ was 0.9, $\boldsymbol{t_0}$ was 100, when $\boldsymbol{t}$ belonged to $(0, \boldsymbol{t_0})$, the range of the value of $\boldsymbol{\lambda}$ was [0.4, 0.9], and the strategy function was a concave function, so the $\boldsymbol{\lambda}$ was growing at a slow speed. Because at the beginning stage, the algorithm wanted to ensure the initial global search, the increase speed trend of damping factor should be slow; when belonged to $(\boldsymbol{t_0}, \boldsymbol{t}_m)$, the range of the value of $\boldsymbol{\lambda}$ was (0.9, 1), from the Fig. 5, the increase speed trend of damping factor was fast, it ensured local search to avoid concussion in the final stage.

And from Fig. 4 to 5, we could find the strategy function change the function shape with the change of time to carry on the parameter search process of damping factor adaptively. According to the different convergence demand of the algorithm in different stages, the strategy function could obtain the suitable damping factor value to control the iteration of ***responsibility*** and ***availability*** in order to achieve the better clustering results.

Thus the section A and section B were the core theories of DPWSAP.

IV. THE ANALYSIS OF SIMULATION EXPERIMENT RESULTS

A. SIMULATION EXPERIMENT

The experiment environment is Pentium G645 2.9 GHz CPU, the memory has 4GB, using MATLAB to implement all codes. The experimental data all use the UCI standard data sets.

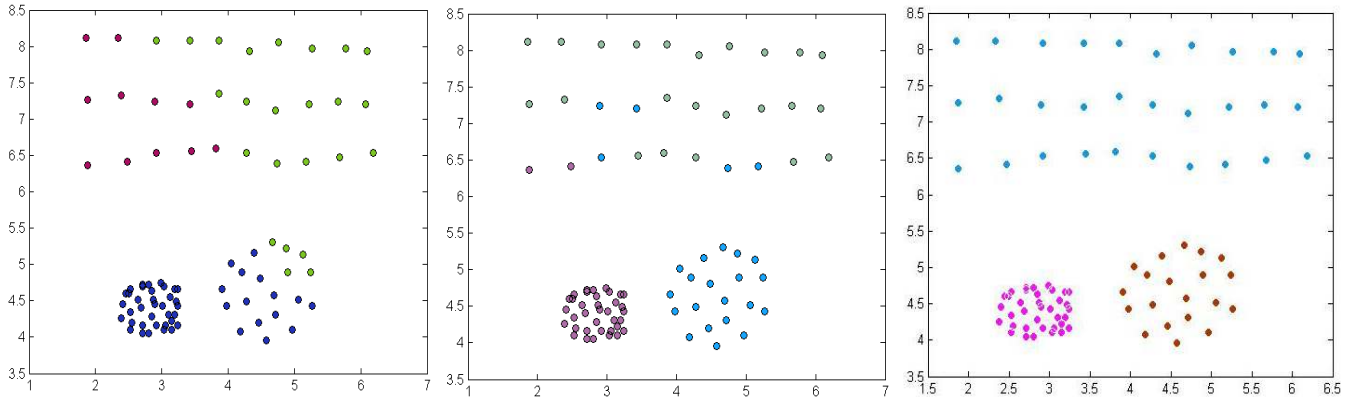
In order to verify the feasibility and effectiveness of DPWSAP algorithm, based on the 6 different UCI data sets, the simulation experiments were carried out, and 4 evaluation indexes were used as the evaluation criteria of clustering quality, the data set was shown in Table 2:

TABLE 2. 6 Different data sets of UCI.

Data Set	Sample Number	Dimension	Class Number
D1	87	2	3
D2	85	2	4
Jain	373	2	2
Iris	150	4	3
Flame	240	2	2
Aggregation	788	2	7

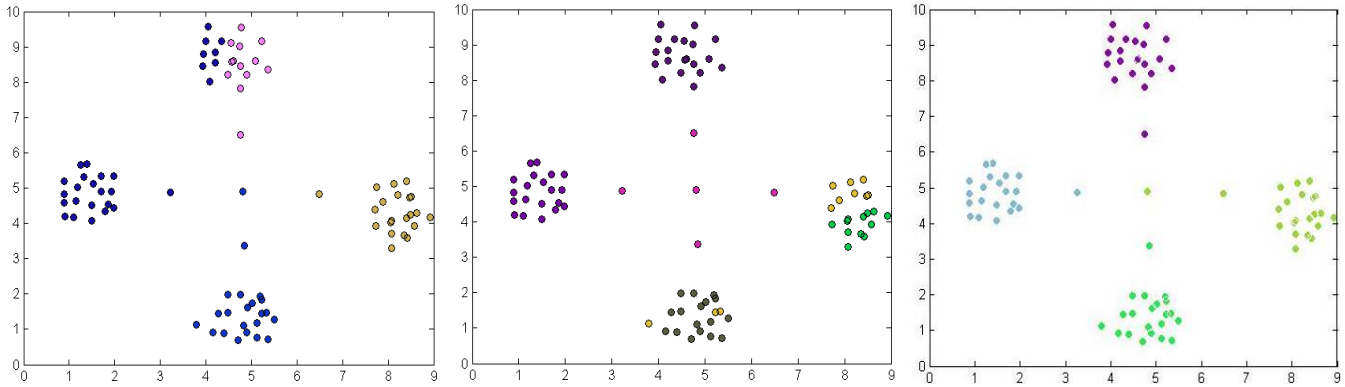
For proving the clustering accuracy of the developed DPWSAP algorithm, this paper selected the two different algorithms that K-means and AP algorithm to compare the clustering results and evaluation results with the DPWSAP algorithm. And we could use the clustering result to reflect the advantage of the DPWSAP algorithm. The simulation experiment of the K-means algorithm, original AP algorithm and DPWSAP algorithm respectively in 6 different data sets were tested, compared the three algorithms of clustering results, and randomly selected 10% data sets as a priori pairwise constraints, K-means algorithm, AP algorithm based on clustering and comparison of DPWSAP algorithm results as shown below.

The results on synthetic data sets were shown in Fig. 6-11, with the 6 data sets, only DPWSAP found the true



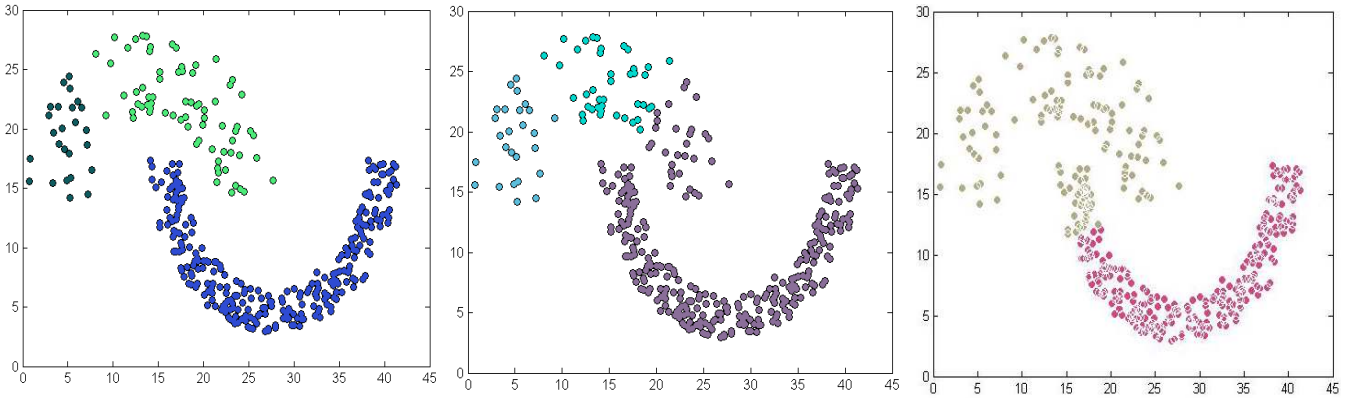
(a) D1 data set with K-means algorithm (b) D1 data set with AP algorithm (c) D1 data set with DPWSAP algorithm

FIGURE 6. **The clustering results image of D1 data set with 3 different algorithms.**



(a) D2 data set with K-means algorithm (b) D2 data set with AP algorithm (c) D2 data set with DPWSAP algorithm

FIGURE 7. **The clustering results image of D2 data set with 3 different algorithms.**

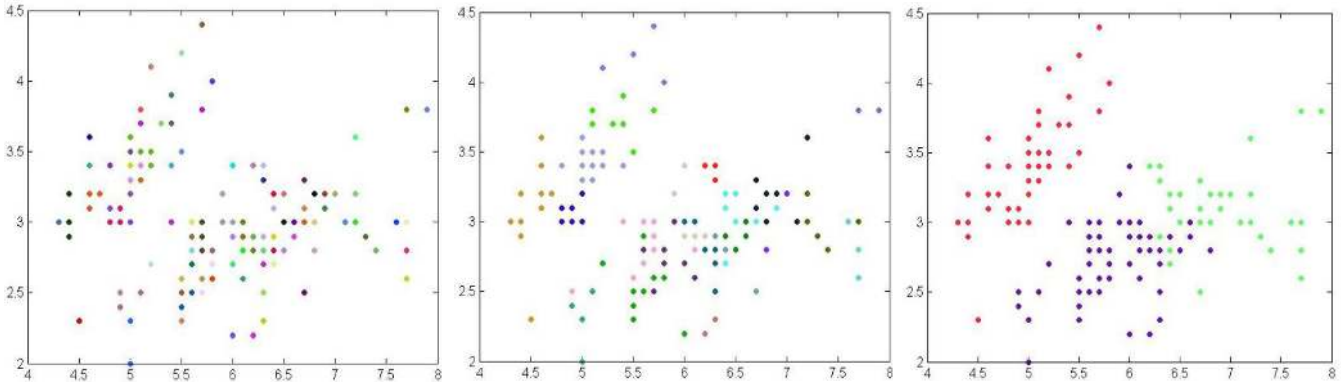


(a) Jain data set with K-means algorithm (b) Jain data set with AP algorithm (c) Jain data set with DAWSAP algorithm

FIGURE 8. **The clustering results image of Jain data set with 3 different algorithms.**

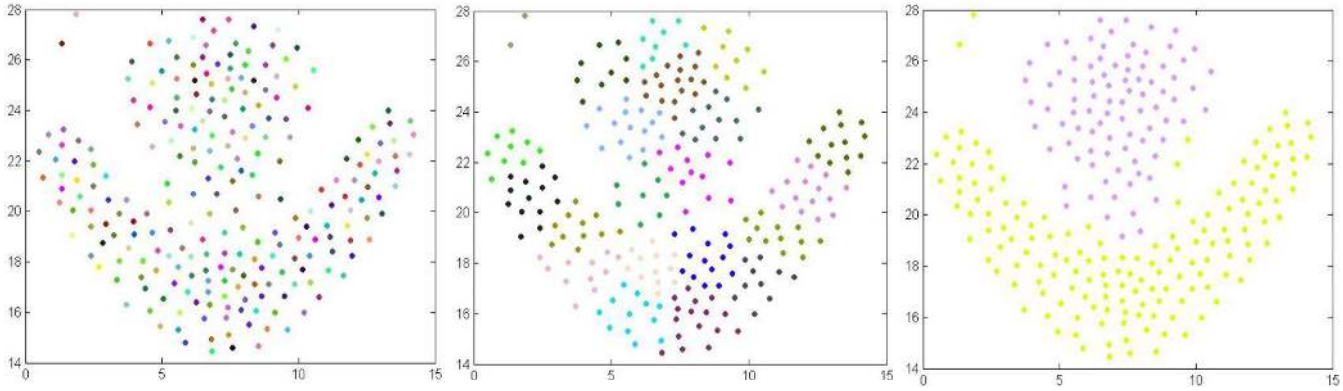
clustering number. AP and K-means obtained similar results on these data sets. However, the three algorithms all cannot find the most accurate clustering result in high-density data set as aggregation.

The number of cluster information was shown in Table 3. And this paper we used the four different external evaluation indicators, including Jaccard, Rand, FM and F1 evaluation indicators.



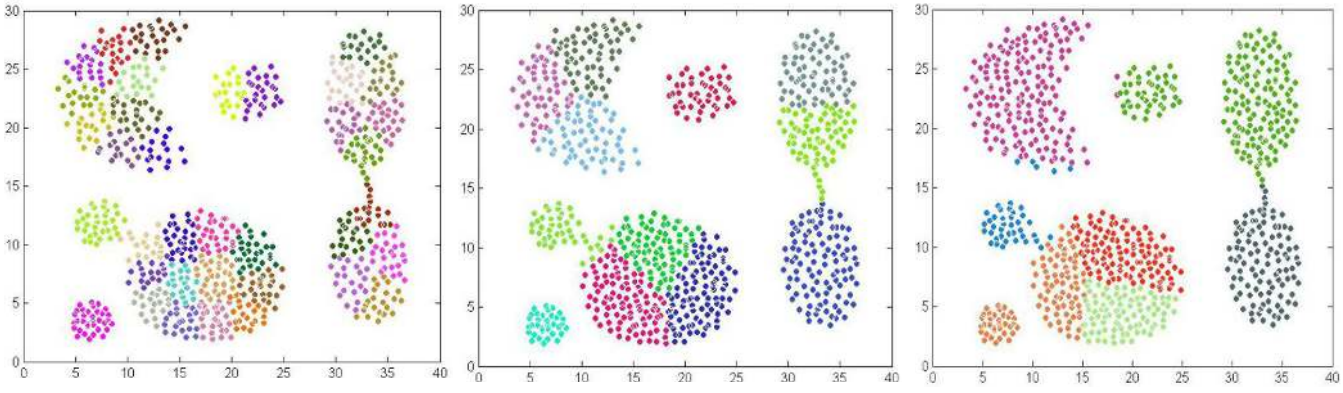
(a) Iris data set with K-means algorithm (b) Iris data set with AP algorithm (c) Iris data set with DPWSAP algorithm

FIGURE 9. The clustering results image of Iris data set with 3 different algorithms.



(a) Flame data set with K-means algorithm (b) Flame data set with AP algorithm (c) Flame data set with DPWSAP algorithm

FIGURE 10. The clustering results image of Flame data set with 3 different algorithms.



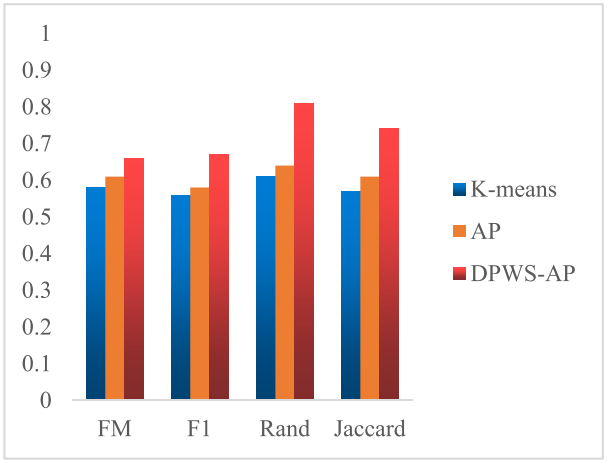
(a) Aggregation data set with K-means algorithm (b) Aggregation data set with AP algorithm (c) Aggregation data set with DPWSAP algorithm

FIGURE 11. The clustering results image of Aggregation data set with 3 different algorithms.

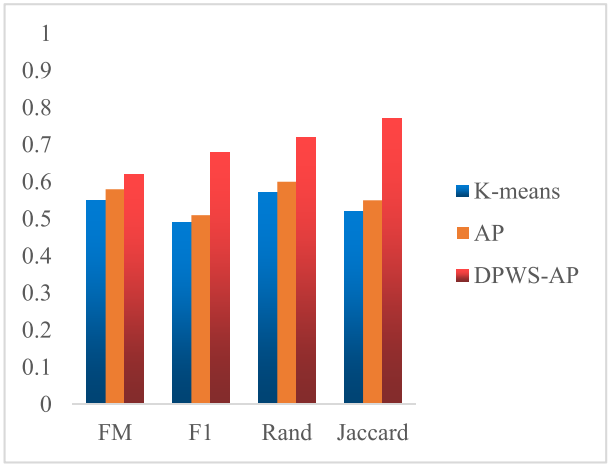
From the table above, we could get that the clustering performance of DPWSAP algorithm was more accurate. Next, this paper would analyze the developed algorithm with the four different evaluating indicators.

According to the number of clusters and the correct clustering results of the known data sets, and make the results of the clustering algorithm named Q to compare with the prior known structure named P, the process is called external evaluation method. For two entities p and q in data set, there are four relationships in P and Q [28].

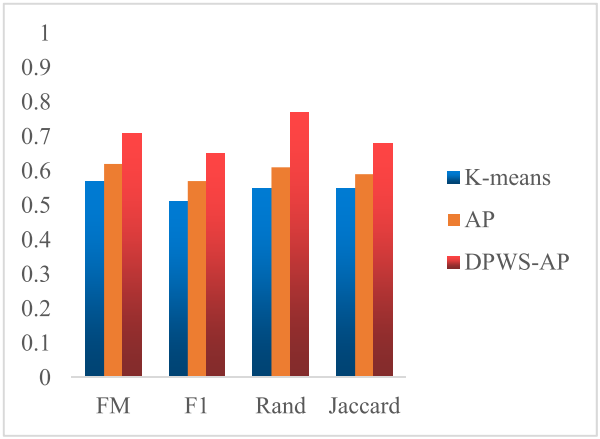
(1) p and q belong to the same class in Q, and belong to the same division in P [28].



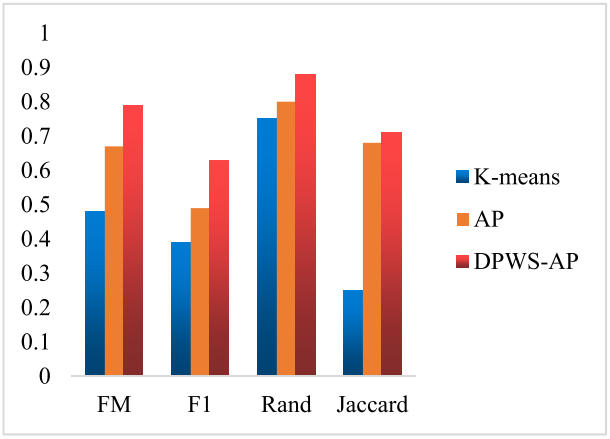
(a) Evaluation index comparison results of D1



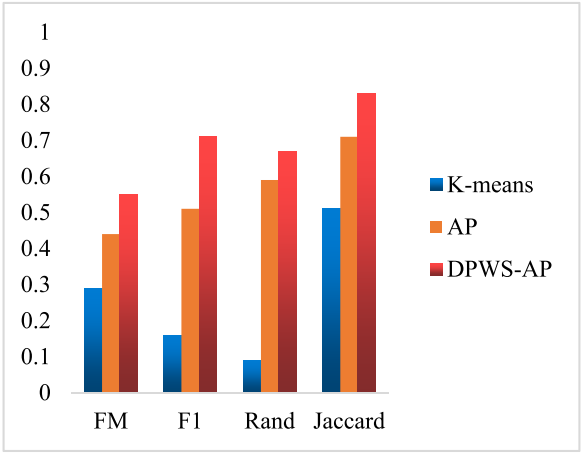
(b) Evaluation index comparison of D2



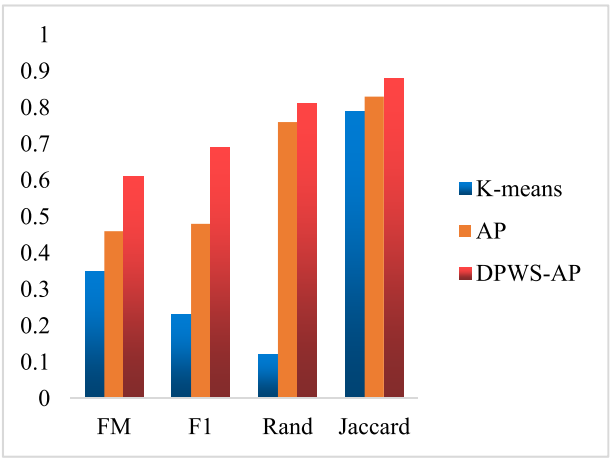
(c) Evaluation index comparison results of Jain



(d) Evaluation index comparison of Iris



(e) Evaluation index comparison of Flame



(f) Evaluation index comparison of Aggregation

FIGURE 12. **The evaluation index results comparison image of 4 data set in 3 different algorithms.**

(2) p and q belong to the same class in Q, but they don't belong to the same division in P [28].
(3) p and q don't belong to the same class in Q, but they belong to the same division in P [28].
(4) p and q don't belong to the same class in Q, and they don't belong to the same division in P [28].

Supposing a, b, c and d satisfy the physical logarithm of the above 4 cases, and M is the sum of the physical logarithm

TABLE 3. Comparison of the clustering number.

Data Set	Class Number	K-means	AP	DPWSAAP
D1	3	4	4	**3**
D2	4	5	5	**4**
Jain	2	3	2	**2**
Iris	3	67	12	**3**
Flame	2	14	12	**2**
Aggregation	7	15	10	**7**

of the data set, and the following relations exist [28].

$$M = a + b + c + d = \frac{N(N-1)}{2} \tag{21}$$

In this formula, the N is the number of entities in the data. According to the above definition, we can attain the formula of the four different evaluation indicators [28].

(1) Jaccard coefficient

$$J = \frac{a}{a+b+c} \tag{22}$$

(2) Rand index

$$R = \frac{a+b}{M} \tag{23}$$

(3) FM index

$$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} \tag{24}$$

(4) F1 index

It combines the idea of recall and precision in the information retrieval domain to cluster evaluation. And exist the formulas [28]:

$$P = precision(i, j) = \frac{N_{ij}}{N_i} \tag{25}$$

$$R = recall(i, j) = \frac{N_{ij}}{N_j} \tag{26}$$

Among them, $\boldsymbol{N_{ij}}$ represents the number of classified $\boldsymbol{i}$ in cluster $\boldsymbol{j}$; $\boldsymbol{N_j}$ represents the number of cluster $\boldsymbol{j}$; $\boldsymbol{N_i}$ represents the number of classified $\boldsymbol{i}$ [28].

$$F1 = \frac{2PR}{P+R} \tag{27}$$

According to these evaluation indicator formulas, this paper objectively compared the three algorithms and obtained that the DPWSAP algorithm was better than the two selected algorithms with the four evaluation indicators. The validity of the algorithm was evaluated as shown in the following figures. From these three evaluation result tables, we can get the DPWSAP algorithm can cluster data more accurate than the K-means algorithm and the original AP algorithm, and the degree of similarity between a class of internal data points is higher, and the degree of similarity between different classes is lower.

B. THE ANALYSIS OF EXPERIMENT RESULTS

Combining with the clustering result figure, we can clearly know that the improved algorithm can not achieve the most accurate clustering results for different types of data. The algorithm can only identify the true class number of some data sets, and can not show the specific classification. This is the aspect of the proposed algorithm to be further improved.

The simulation results from Fig. 12 showed that the DPWSAP algorithm clustered the data on the basis of the above six data sets, which fully consistent with the data set for real class number. It explained that the DPWSAP algorithm could carry on reasonable data clustering in order to achieve the real clustering requirements, and the clustering effect was better [29], [30].

V. CONCLUSION

This paper proposed DPWSAP to solve both the similarity calculation method and parameter limits problems of the affinity propagation algorithm. Different from the existing clustering algorithms, DPWSAP put forward the fusion of domain density and distance method to calculate the similarity more accurately. Instead of using a single computing approach, also the growth curve function model was introduced to enhance the convergence ability of the damping factor. The model was a self-adaptive process, which enhance the convergence performance of AP at different stages when aiming at different searching requirements.

Extensive comparative studies were conducted based on 6 different test data in 3 different clustering algorithms. The experimental results demonstrated that the proposed DPWSAP was an efficient clustering algorithm. As a whole, DPWSAP performed better than other involved algorithms in solving these 6 different UCI data sets with different density, different dimensions, and different data volumes. This paper applied DPWSAP to analyze the Chinese economic situation with economic indicator data. The analysis result showed that the DPWSAP had a good application value. Besides, the proposed method still has some limitations for data sparsity and it is the future direction of work [31]–[37].

REFERENCES

[1] S. Ayoubi, N. Limam, and M. A. Salahuddin, N. Shahriar, R. Boutaba, F. Estrada-Solano, and O. M. Caicedo, "Machine learning for cognitive network management," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 158–165, Jan. 2018.

[2] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[3] F. T. C. Tan, Z. Guo, M. Cahalane, and D. Cheng, "Developing business analytic capabilities for combating e-commerce identity fraud: A study of Trustev's digital verification solution," *Inf. Manage.*, vol. 53, no. 7, pp. 878–891, Nov. 2016.

[4] C. Lee and H. Kim, "The evolutionary trajectory of an ICT ecosystem: A network analysis based on media users' data," *Inf. Manage.*, vol. 55, no. 6, pp. 795–805, Sep. 2018.
[5] R. Atat, L. Liu, J. Wu, G. Li, C. Ye, and Y. Yang, "Big data meet cyber-physical systems: A panoramic survey," *IEEE Access*, vol. 6, pp. 73603–73636, 2018.
[6] W. Serrano, "Neural networks in big data and Web search," *Data*, vol. 4, no. 1, p. 7, Dec. 2018.
[7] S. J. Miah, H. Q. Vu, J. Gammack, and M. McGrath, "A big data analytics method for tourist behaviour analysis," *Inf. Manage.*, vol. 54, no. 6, pp. 771–785, Sep. 2017.
[8] Y. Wang, L. A. Kung, W. Y. C. Wang, and C. G. Cegielski, "Integrated big data analytics-enabled transformation model: Application to health care," *Inf. Manage.*, vol. 55, no. 1, pp. 64–79, Jan. 2018.
[9] J. Wu, H. Li, S. Cheng, and Z. Lin, "The promising future of healthcare services: When big data analytics meets wearable technology," *Inf. Manage.*, vol. 53, no. 8, pp. 1020–1033, Dec. 2016.
[10] Y. Fujiwara, G. Irie, and T. Kitahara, "Fast algorithm for affinity propagation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Jan. 2011, pp. 1–6.
[11] B. Jia, B. Yu, Q. Wu, C. Wei, and R. Law, "Adaptive affinity propagation method based on improved cuckoo search," *Knowl.-Based Syst.*, vol. 111, pp. 27–35, Nov. 2016.
[12] X. Feng and H. Yu, "Semi-supervised affinity propagation clustering based on manifold distance," *Appl. Res. Comput.*, vol. 28, no. 10, pp. 3656–3658, 2011.
[13] 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 Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1152–1156, Sep. 2013.
[14] 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, Jul. 2019.
[15] L. Wang, Q. Ji, and X. Han, "Adaptive semi-supervised affinity propagation clustering algorithm based on structural similarity," *Tech. Gazette*, vol. 23, no. 2, pp. 425–435, Apr. 2016.
[16] A. Valls, K. Gibert, A. Orellana, and S. Antón-Clavé, "Using ontology-based clustering to understand the push and pull factors for British tourists visiting a mediterranean coastal destination," *Inf. Manage.*, vol. 55, no. 2, pp. 145–159, Mar. 2018.
[17] Z. Wei, Y. Wang, and S. He, "A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection," *Knowl.-Based Syst.*, vol. 116, pp. 1–12, Jan. 2017.
[18] L. Wang, X. Zhou, and Y. Xing, "Clustering ecg heartbeat using improved semi-supervised affinity propagation," *IET Softw.*, vol. 11, no. 5, pp. 207–213, Oct. 2017.
[19] P. Sadhukhan, "Performance analysis of clustering-based fingerprinting localization systems," *Wireless Netw.*, vol. 51, pp. 1–14, Feb. 2018.
[20] H. Wang, R. Nie, X. Liu, and T. Li, "Constraint projections for semi-supervised affinity propagation," *Knowl.-Based Syst.*, vol. 36, pp. 315–321, Dec. 2012.
[21] B. Hassanabadi, C. Shea, L. Zhang, and S. Valaee, "Clustering in vehicular ad hoc networks using affinity propagation," *Ad Hoc Netw.*, vol. 13, pp. 535–548, Feb. 2014.
[22] F. Shang, L. C. Jiao, J. Shi, F. Wang, and M. Gong, "Fast affinity propagation clustering: A multilevel approach," *Pattern Recognit.*, vol. 45, no. 1, pp. 474–486, 2012.
[23] P. Li, H. Ji, B. Wang, Z. Huang, and H. Li, "Adjustable preference affinity propagation clustering," *Pattern Recognit. Lett.*, vol. 85, pp. 72–78, Jan. 2017.
[24] J. Meng, H. Hao, and Y. Luan, "Classifier ensemble selection based on affinity propagation clustering," *J. Biomed. Inform.*, vol. 60, Apr. 2016, pp. 234–242.
[25] R. Ma, Q. Guo, C. Hu, and J. Xue, "An improved WiFi indoor positioning algorithm by weighted fusion," *Sensors*, vol. 15, no. 9, Aug. 2015, Art. no. 2182421843.
[26] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[27] 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," *Knowl.-Based Syst.*, vol. 133, pp. 294–313, Oct. 2017.
[28] L. Wang and Z. Hao, "Gravity theory-based affinity propagation clustering algorithm and its applications," *Tech. Gazette*, vol. 25, no. 4, pp. 1125–1135, Apr. 2018.
[29] W. Zhang, X. Wu, W.-P. Zhu, and L. Yu, "Unsupervized image clustering with SIFT-based soft-matching affinity propagation," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 461–464, Apr. 2017.
[30] K. Ebner, N. Urbach, and B. Mueller, "Exploring the path to success: A review of the strategic IT benchmarking literature," *Inf. Manage.*, vol. 53, no. 4, pp. 447–466, Jun. 2016.
[31] 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," *Appl. Intell.*, vol. 49, no. 4, pp. 1245–1259, Apr. 2019.
[32] J. Bi, Y. Wang, X. Li, H. Qi, H. Cao, and S. Xu, "An adaptive weighted KNN positioning method based on omnidirectional fingerprint database and twice affinity propagation clustering," *Sensors*, vol. 18, no. 8, p. 2502, Aug. 2018.
[33] Y. Kokkinos and K. G. Marfaritis, "Kernel averaged gradient descent subtractive clustering for exemplar selection," *Evol. Syst.*, vol. 9, no. 4, pp. 285–297, Dec. 2018.
[34] P. A. Karegar, "Wireless fingerprinting indoor positioning using affinity propagation clustering methods," *Wireless Netw.*, vol. 24, no. 8, pp. 2825–2833, Nov. 2018.
[35] E. Graham, J. Heidelberg, and B. Tully, "BinSanity: Unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation," *Peerj*, vol. 5, p. e3035, Mar. 2017.
[36] D. Chen, J. Sheng, and J. Chen, "Stability-based preference selection in affinity propagation," *Neural Comput. Appl.*, vol. 25, nos. 7–8, pp. 1809–1822, Dec. 2014.
[37] L. Yaohui, M. Zhengming, and Y. Fang, "Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy," *Knowl.-Based Syst.*, vol. 133, pp. 208–220, Oct. 2017.

LIMIN WANG received the master's and Ph.D. degrees in computer science and technology from Jilin University, in 2004 and 2007, respectively. She is currently a Professor. Her current research interests include big data analysis, evolutionary algorithm, and intelligent decision optimization. She has authored or coauthored more than 81 research articles in international or domestic journals or international conference, and some of them have been indexed by SCI or EI. In recent years, she has presided or completed more than 40 research projects. She is a member of the China Computer Federation.

ZHIYUAN HAO received the master's degree from the Jilin University of Finance and Economics, in 2016, and the master's degree in management science and engineering. He is currently pursuing the Ph.D. degree with the School of Management, Jilin University, China. His research interests include computer science, big data analysis, data mining, machine learning, and intelligent decision optimization.

WENJING SUN received the bachelor's degree in computer science and technology, in 2014 and master's degree in library and intelligence. She is currently pursuing the master's degree with the School of Management Science and Information Engineering, Jilin University of Finance and Economics, China. Her research interests include big data analysis, data mining, and swarm intelligent optimization.