ISSN 1330-3651(Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20180504150204
Original scientific paper

Gravity Theory-Based Affinity Propagation Clustering Algorithm and Its Applications

Limin WANG, Zhiyuan HAO, Xuming HAN, Ruihong ZHOU

Abstract: The original Affinity Propagation clustering algorithm (AP) only used the Euclidean distance of data sample as the only standard for similarity calculation. This method of calculation had great limitations for data with high dimension and sparsity when the original algorithm was running. Due to the single calculation method of similarity, the convergence and clustering accuracy of the algorithm were greatly affected. On the other hand, in the universe, we can consider the formation of galaxies is a clustering process. In addition, the interaction between different celestial bodies are achieved through universal gravitation. This paper introduced the Density Peak clustering algorithm (DP) and gravitational thought into the AP algorithm, and constructed the density property to calculate the similarity, put forward the Affinity Propagation clustering algorithm based on Gravity (GAP). The proposed algorithm was more accurate to calculate similarity of simple points through the local density of corresponding points, and then used the gravity formula to update the similarity matrix. The data clustering process could be seen as the sample points spontaneously attract each other based on 'gravitation'. Experimental results showed that the convergence performance of GAP algorithm is obviously improved over the AP algorithm, and the clustering effect was better.

Keywords: affinity propagation algorithm; gravitation theory; local density; similarity matrix

1 INTRODUCTION

With the popularization and application of internet technology, more and more data appear in electronic form and grow in the form of explosion. In the information age, all occupations have gradually begun to use computer technology for data management, greatly improving the ability to generate, collect, store and process data. However, the database system analysis tool with more limitations, and extracted from the database of valuable knowledge accounted for only a small part of the whole database system of knowledge, which leads to "more data, less knowledge" phenomenon. How to avoid the waste of data resources, how to extract more and more valuable information from the mass of data and further improve the utilization of data, etc. These force us to find new and more effective data analysis tools, data mining appeared in this context [1]. At the same time, with the arrival of the era of big data and the gradual maturity of artificial intelligence, human civilization has entered a new era of data intensive, data had become a very important asset. Thanks to the rapid development of data acquisition and storage technology, large databases and server systems are accessing data, accessing data and statistics in a high efficiency and low energy consumption way [2]. However, in the face of all the massive data continue to surge, the real human can obtain valuable information is scanty from it, we gradually found that the mass of information had not brought convenience to people, but also brings difficulties and challenges hitherto unknown. In view of this, how to effectively analyse and use the vast amounts of raw data and dig out valuable information had become a hot topic of many researchers [3]. Clustering analysis is an important branch of data mining. It analyses data without any prior information and divides the data into meaningful or useful arrays, also called unsupervised learning [4]. So far, clustering has been widely used in many fields such as pattern recognition, data analysis, image processing, market research, facility location, and so on [5]. At present, the clustering algorithm mainly includes hierarchical clustering algorithm, clustering algorithm based on partition and based on density and grid clustering algorithm. For example, we all know the K-means clustering algorithm, it is the most widely used clustering algorithm based on partitioning, the algorithm is simple and efficient, and so it is widely used. However, K-means clustering algorithm is a greedy algorithm, easy to fall into the local extremum problem and the clustering results are limited by the selection of the initial class representative points, so many scholars are constantly looking for new clustering algorithm.

Affinity propagation is a kind of reasonable and accurate clustering algorithm that based on similarity degree matrix for the construction [6]. Compared with the traditional K-means algorithm, etc. affinity algorithm has the advantage that it does not need to be sure the clustering centre, the all test data points clustering centre has become a possibility, and also it can iterate constantly data samples with process of running by responsibility and the availability to get the optimal clustering centre [7, 8]. Being based on the characteristic, affinity algorithm has been widely used in various kinds of field data clustering analysis. At the current, the all over world scholars have made many improvements and continued a lot of research in the affinity propagation algorithm. In the literature [9], Fujiwara and other writers in order to promote the convergence speed of affinity propagation algorithm, and on the premise of ensure accuracy of clustering, they delete unnecessary information in the process of operation of the algorithm [10, 11]. Moreover, in the literature [12], based on 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 that based on the singular value decomposition, and the improvement can better solve the problem about the high dimensional data of affinity propagation algorithm to further improve the clustering effect of the algorithm.

Although the AP algorithm has many incomparable advantages over the other algorithms, with good and stable effect and in practical application, but it still faces some difficulties and challenges: (1) although the AP algorithm does not need to specify the number of clusters can get optimal cluster centre competition by information exchange, but the algorithm is biased the parameters of P is not set based on the clustering structure of data itself, so only by manual adjustment to find the optimal clustering results, it will increase the cost of using the algorithm. (2) due to the traditional AP algorithm uses the Euclidean distance similarity between samples, algorithm in dealing with high dimensional data clustering problem, due to the high dimensional data sparsity and space characteristics, the Euclidean distance can accurately and reasonably reflect the similarity between data objects, this algorithm will directly lead to failure. (3) the traditional AP algorithm is based on Euclidean distance as the similarity measure for clustering, so in dealing with cluster structure similar to spherical or ellipsoidal shape has a good clustering effect, and in the face of cluster structure more complex data, can effectively identify the true clustering potential structure of data, resulting in limited the application scope of the algorithm. In view of this, based on the traditional AP algorithm advantages, effectively improve the shortcomings, improve the performance of the algorithm, which makes it more widely applied in the practical work, for the government and enterprises to provide a more effective basis for decision-making, has a very important significance [14, 15].

In the process of the original affinity propagation algorithm calculates similarity. The original Affinity Propagation clustering algorithm only used the Euclidean distance of data sample as the only standard for similarity calculation. This method of calculation had great limitations for data with high dimension and sparsity when the original algorithm was running. Due to the single calculation of similarity, the convergence and clustering accuracy of the algorithm were greatly affected. In the universe, we can consider the formation of galaxies is a clustering process, the interaction between different celestial bodies is achieved through universal gravitation. This paper introduced the Density Peak clustering algorithm and gravitational thought into the AP algorithm, and constructing the density property to calculate the similarity, put forward the Affinity Propagation clustering algorithm based on Gravity. The proposed algorithm was more accurately to calculate similarity of simple points through the local density of the corresponding points, then, used the gravity formula to update the similarity matrix. The data clustering process could be seen as the sample points spontaneously attract each other based on 'gravitation'. Experimental results shows that the convergence performance of GAP algorithm is obviously improved, and the clustering effect is better.

2 AFFINITY PROPAGATION ALGORITHM

Affinity propagation algorithm called AP algorithm. It is a more popular clustering algorithm in recent years. AP algorithm mainly carry on clustering research that based on the degree of similarity between data objects. When the algorithm carry on clustering, the algorithm does not need to specify the clustering centre in advance, and it means that the AP algorithm can discretionarily assign any data as the necessary clustering centre in the data sample. At the same time, the degree of similarity between data objects can be symmetrical or asymmetrical. And the advantages of the affinity propagation algorithm have lay in that it can count each data point as a centre of clustering and it can be performed, also, it don't have any limit of the clustering number. As well as, the affinity propagation algorithm by reference to the degree of P to estimate the feasibility for an arbitrary point as the clustering centre [16]. Accordingly, the larger select the P value, the data point is the more possible to become the clustering centre, at the same time, it can obtain a clustering with the more class number. If the value is relatively smaller, the result can account for the data point to be a clustering centre in the data sample with the less possibility, of course, correspondingly, it can exist the clustering with the less class.

In the high-dimensional space, freely choose two data objects (for example, can be selected as s and t) to carry on the clustering analysis, finally, using data obtained by the AP algorithm is mainly based on the following two types: responsibility data and availability data. The responsibility data mainly reflects that in the data set (*i*, *k*), the *k* point set can be the appropriateness degree for the clustering centre of data point set *i*. In a similar way, availability data mainly reflects that in the set (*i*, *k*), *i* choose the possibility of *k* as a core size. In the course of the iteration, the algorithm searches the clustering centre until find enough suitable clustering centre [17].

The AP algorithm has the following characteristics:

1) Compared with the traditional clustering algorithm, the affinity propagation algorithm without any restrictions for the specified object and parameters [18].
2) Affinity propagation algorithm does not need to carry on the selection of initial value [19].
3) If use the error sum of squares to be the standard to compare the advantages and disadvantages between various algorithm. The error sum of squares of affinity propagation algorithm is the lowest compared with other algorithms [20].
4) AP algorithm using the correlation matrix to start, so no mandatory requirement for the distribution of the data.
5) AP algorithm's time complexity is high, when the AP algorithm in complex clustering analysis, the algorithm running time will be longer.

In affinity propagation algorithm, for any two points in the sample space between x_i and x_k, the similarity expressed in $s(i, k)$. The mathematical expression are the following:

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

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

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

The core of AP is 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 *r*(:, *k*)+*a*(:, *k*), more probability the point k as a final cluster centre. A decision matrix E is calculated after each update. Decision matrix E represents whether point i chooses point k as its exemplar or not. The following formulas completely reflect the basic process of AP algorithm [21, 22].

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

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

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

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

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

3 ALGORITHM DESCRIPTION

3.1 The Basic Principle of Gravitation Theory

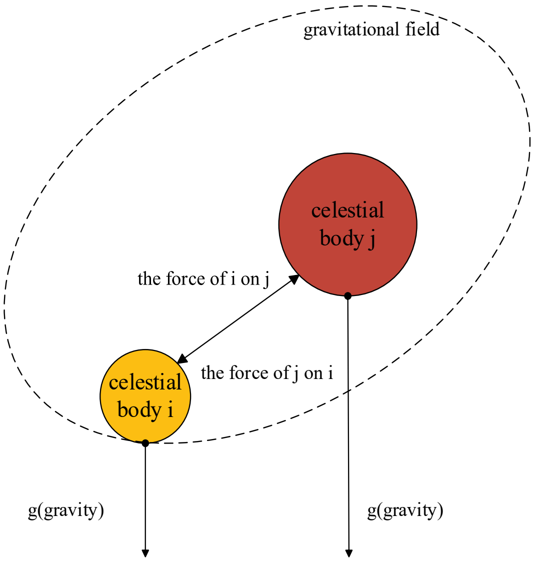

Figure 1 Celestial bodies interaction

The law of gravity unifies the laws of motion of objects on the earth and the celestial bodies, on the other way, the formation of galaxies in the universe and the clustering analysis of data sets are very similar, thus, each galaxy can be regarded as a class of cluster analysis. In the early days of the universe, various substances are randomly distributed in every corner of the universe, because of the existence of gravitational attraction, and therefore it cause two gravitational masses to cluster together and evolve into galaxies [23, 24]. Any two objects in nature are attracted to each other, the magnitude of gravity is directly proportional to the product of their mass, and they are inversely proportional to the square of their distance. The interaction relation between any two objects is expressed by the following formula.

$$F=G\cdot\frac{M\cdot m}{r^{2}} \tag{8}$$

In the formula, G is the acceleration of gravity, and the M and the m are the quality attributes of corresponding objects, r is the distance between two objects.

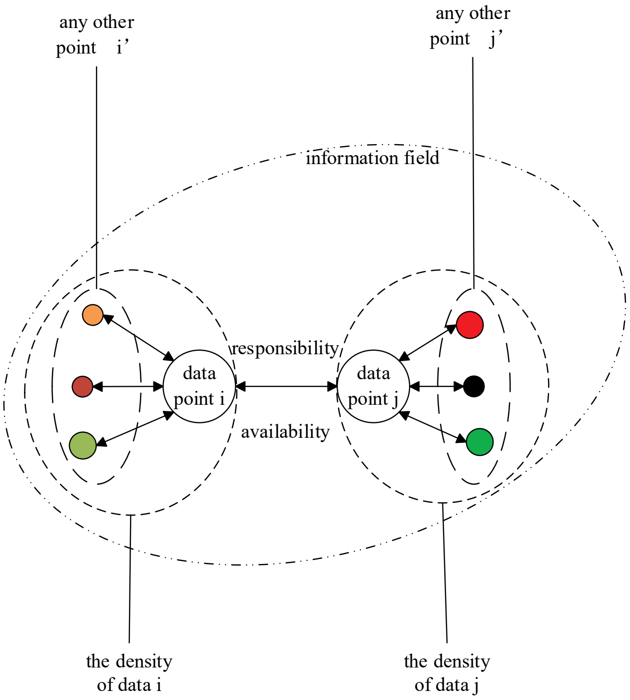

Figure 2 Information transfer between data points

All data in the sample space can be regarded as a similar object, with introducing the basic idea of gravity, it can be assumed that any of the two data points in the sample data has attractive properties that attract each other. When the data point i' is infinitely closed to the point 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 DPC algorithm, and in the DPC algorithm, the local density depended on the cut-off distance, and the density of the data point i can be calculated by the number of points in a range, and cut-off distance is smaller, and the points in the range are more closed to the point i. Considering that the basic property of the quality of data points in the sample space is non-existent, this paper uses the local density of the density peak clustering algorithm, the local density is introduced into the original affinity propagation algorithm, and the local density of data points is substituted for the quality attributes. This paper uses gravity to update similarity matrix, and a new variable local density is added on the basis of the original algorithm, from the two aspects of the local density and the distance which between the any two data points, we can better describe the similarity, as a result, the convergence performance of the algorithm is improved and the clustering results are enhanced. The improved formula for gravity is as follows:

$$S=G\cdot\frac{\rho_{1}\cdot\rho_{2}}{r^{2}} \tag{9}$$

In the formula, G is the acceleration of gravity, ρ_1 and ρ_2 are the local density of the data points, and r is the distance between two points.

3.2 Local Density

There are two important rules in the density peak clustering algorithm. The one is that the density of a defined cluster class centre is higher than that of other adjacent points. In addition, another one is that the defined cluster class centres are relatively far away from points with higher density values. The original Density Peak Cluster (DPC) algorithm uses Euclidean distance to calculate the distance between the any two data points [25]. Assume the presence of data points x_i and x_j, then the distance between them is the following:

$$d_{ij} = distance\left(x_i, x_j\right) \tag{10}$$

For each data point x_i has a density value of ρ_i, the formula is the following:

$$\rho_i = \sum_j \chi \cdot \left(d_{ij} - d_c\right) \tag{11}$$

The $\chi(i)$ is a function, when the $x < 0$, the $\chi = 1$, otherwise, the $\chi = 0$, the d_c in the formula is a parameter of the DPC algorithm that named Cut-off distance. The selection of the Cut-off distance d_c affects the results of the whole DPC clustering, in the original DPC algorithm, this parameter is determined by the distance from the first 1-2% in the distance from the data point in ascending order. In Eq. (6), the calculation method of medium density value is similar to the calculation method of MinPoints in DBSCAN algorithm, that they are all define the number of correlation points in the neighbourhood of the cluster centre. In other way, ρ_i can also be expressed that use Eq. (12).

$$\rho_i = \sum_j \exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \tag{12}$$

3.3 Affinity Propagation Clustering Algorithm Based on Gravity (GAP)

3.3.1 Determine the d_c Value

There are two important parameters in the density peak clustering algorithm. The one is the local density ρ, and the other one is the distance from the nearest point of high density δ. The algorithm constructs the decision diagram through two parameters, to determine the cluster centre and complete the clustering. Moreover, these two important parameters need to be based on the Cut-off distance d_c value. Therefore, determining the proper d_c value is the key to density peak clustering.

Cut-off distance d_c is that sort-ascending data by sorting distances between all any two data points. By experience, a suitable distance is chosen as the value of the Cut-off distance. In the original peak density algorithm, the Cut-off distance is between 1-2%. However, because of the difference of data selection, there is a great difference in the value of the Cut-off distance, therefore, in this paper, the author on the premise of obtaining better clustering results to determine the d_c value [26].

3.3.2 Density Peak Algorithm is used to Calculate Density

Density peak clustering has good clustering centre exploration ability; the reason is that the algorithm is reasonable and convenient to set up the basic parameters. Therefore, in the process of building the structure similarity, the author try to update the similarity matrix by introducing the local density in the peak value [27].

At the same time, in order to better adapt to the clustering of low-density data sets, the similarity distance based on Gauss kernel function is used to guide the generation of local density. The calculation method is as follows:

$$\rho_i = \sum_j \exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \tag{13}$$

3.3.3 Using Gravity to Obtain Similarity Matrix

All data in the sample space can be regarded as a similar object, introducing the basic idea of gravity, it can be assumed that any of the two data points in the sample data has attractive properties that attract each other. Considering that the basic property of the quality of data points in the sample space is non-existent, this paper uses the local density of the density peak clustering algorithm, the local density is introduced into the original affinity propagation algorithm, and the local density of data points is substituted for the quality attributes. The specific formula is as follows:

$$S = G \cdot \frac{\rho_1 \cdot \rho_2}{r^2} \tag{14}$$

The GAP algorithm combined the local density of DPC algorithm and the gravity theory, and this paper used the two theories to change the similarity of the traditional Affinity Propagation Cluster algorithm, and using the gravity formula to set up a new way to calculate the similarity of the data sample. Moreover, to some extent, this development can reduce the adaptability of the original affinity propagation cluster algorithm to high-dimensional data, and it reasonably uses the local density, therefore, the clustering accuracy of the affinity propagation algorithm is improved. At the same time, for the obtaining of the similarity, the GAP algorithm also can run better, and the process of the GAP is following.

The GAP algorithm is shown in Tab. 1.

Table 1 The process of GAP algorithm

Input: Similarity matrix S(i, j), Cut-off distance d_c value
Output: Cluster number k, Division result C={C_1,…, C_k}, F-measure
Step1: Select d_c value
Step2: Density peak algorithm is used to calculate density ρ
Step3: Using gravity to obtain similarity matrix
Step4: Using the similarity matrix to guide the clustering of AP algorithm and to obtain the clustering results
Step5: Run the AP algorithm, and use the F-measure to evaluate the effectiveness of the algorithm
Step6: Record the clustering results

4 THE ANALYSIS OF SIMULATION EXPERIMENT RESULTS

4.1 Simulation Experiment

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

In order to verify the feasibility and effectiveness of GAP algorithm, based on the 4 sets of UCI data sets, the simulation experiments are carried out, and 4 evaluation indexes are used as the evaluation criteria of clustering quality, the data set is shown in Tab. 2.

Table 2 Data sets of UCI

Data Set	Sample Number	Dimension	Class Number
Iris	150	4	3
Wine	178	13	3
Flame	240	2	2
Aggregation	788	2	7

The simulation experiment of the original AP algorithm and GAP algorithm respectively in 4 different data sets were tested, compared two algorithms of clustering results, and randomly selected 10% data sets as a priori pairwise constraints, the comparison result of AP algorithm clustering results and the affinity propagation algorithm of gravity results as shown below. Each of these photos following were the clustering results of AP algorithm and the GAP algorithm. In this paper, we compared the 4 different data sets and got the different clustering results, the following clustering results were drawn, on the other hand, the proposed algorithm were further evaluated by the evaluation index.

The following were the comparison figures of the clustering results for two algorithms.

4.1.1 For Iris Data Set

From Fig. 3 and Fig. 4, for the iris data set, the GAP algorithm can obtain 3 classes obviously, but the traditional AP algorithm attain 12 classes, and we can get the clustering accuracy of the improved algorithm is shown more.

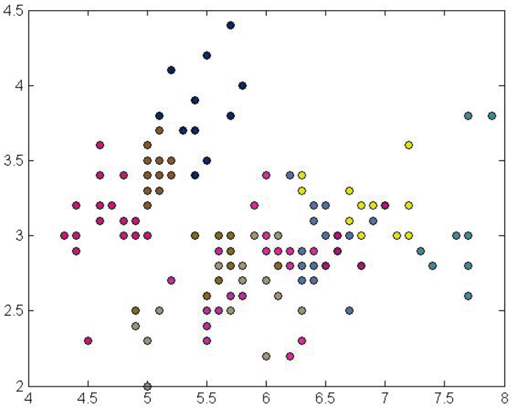

Figure 3 Iris data set with AP algorithm

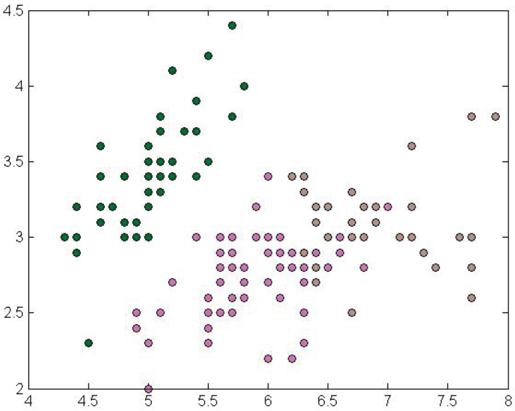

Figure 4 Iris data set with GAP algorithm

4.1.2 For Wine Data Set

From Fig. 5 and Fig. 6, for the wine data set, the GAP algorithm can obtain 3 classes obviously, but the traditional AP algorithm attain 12 classes, and we can get the clustering accuracy of the improved algorithm is shown more.

However, when running the proposed algorithm, the real clustering result was still not very accurate.

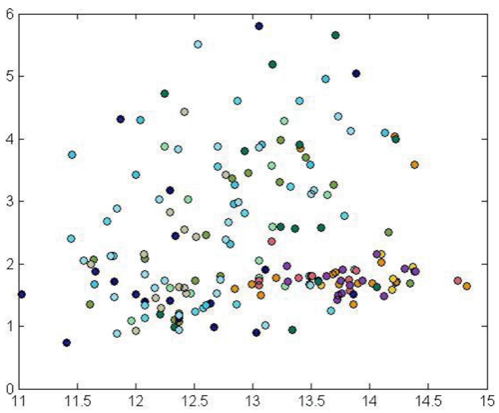

Figure 5 Wine data set with AP algorithm

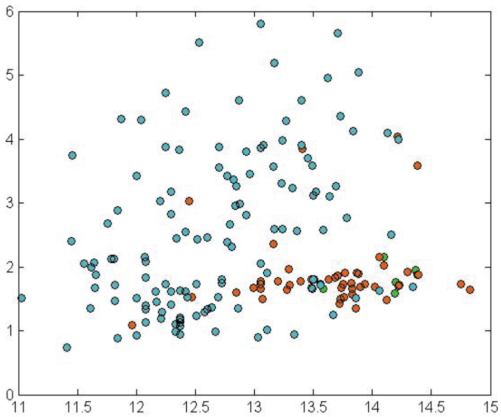

Figure 6 Wine data set with GAP algorithm

4.1.3 For Flame Data Set

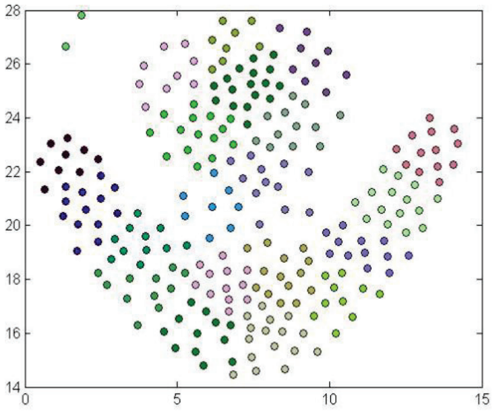

Figure 7 Flame data set with AP algorithm

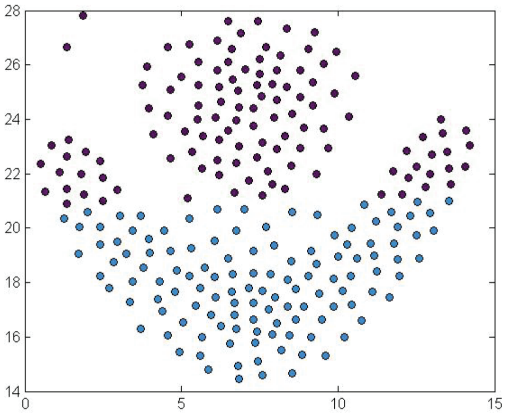

Figure 8 Flame data set with GAP algorithm

4.1.4 For Aggregation Data Set

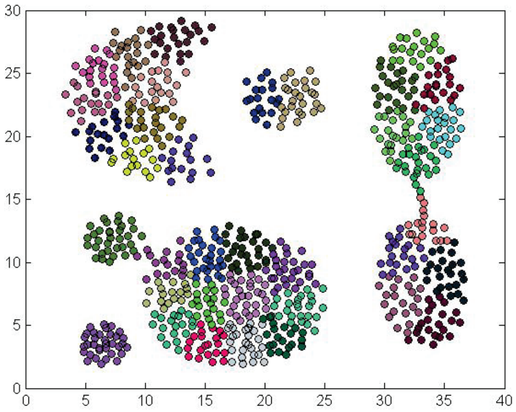

Figure 9 Aggregation data set with AP algorithm

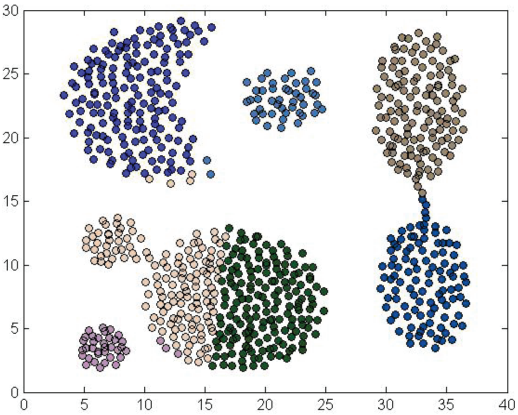

Figure 10 Aggregation data set with GAP algorithm

In order to prove the improved algorithm that the clustering performance has been improved greatly, in this paper, we choose these four data set that the traditional affinity propagation cannot obtain the accurate cluster number. According to the above 8 graphs of clustering results that the improved algorithm compared with the original affinity propagation algorithm has significantly improved the performance of clustering algorithm, and the test data set can obtain more accurate clustering number, of course, for some special type data set, although the GAP can get the accurate clustering number, the developed algorithm cannot get the perfect cluster yet, like the aggregation data set and the flame data set, we can get the accurate cluster number, but can not obtain the perfect cluster result. For the other hand, we can compare the two algorithm in the different evaluation index comparison results, and the following are the concrete comparison results.

The number of cluster information is shown in Tab. 3. Moreover, this paper we use the four different external evaluation indicators, including Jaccard, Rand, FM and F1 evaluation indicators [28].

Table 3 Comparison of the clustering number

Data Set	Class Number	AP	GAP
Iris	3	12	3
Wine	3	12	3
Flame	2	21	2
Aggregation	7	36	7

According to the number of clusters and the correct clustering results of the know data sets, and make the results of the clustering algorithm that named Q compare with the prior known structure that named P, the process which is called external evaluation method. For two entities p and q in data set, there are four relationships in P and Q [29, 30].

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

Supposing a, b, c and d satisfy the physical logarithm of the above 4 cases, and M is the sum of the physical logarithm of the data set, and the following relations exist.

$$M = a + b + c + d = \frac{N(N-1)}{2} \tag{15}$$

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.

1) Jaccard coefficient

$$J = \frac{a}{a+b+c} \tag{16}$$

2) Rand index

$$R = \frac{a+b}{M} \tag{17}$$

3) FM index

$$FM = \sqrt{\frac{\text{a}}{a+b}\frac{a}{a+c}} \tag{18}$$

4) F1 index

It combines the idea of recall and precision in information retrieval domain to cluster evaluation, and exist the formulas:

$$P = precision(i, j) = \frac{N_{ij}}{N_i} \tag{19}$$

$$R = recall(i, j) = \frac{N_{ij}}{N_j} \tag{20}$$

Among them, N_{ij} represents the number of classified i in cluster j; N_j represents the number of cluster j; N_i represents the number of classified i.

$$F1 = \frac{2PR}{P+R} \tag{21}$$

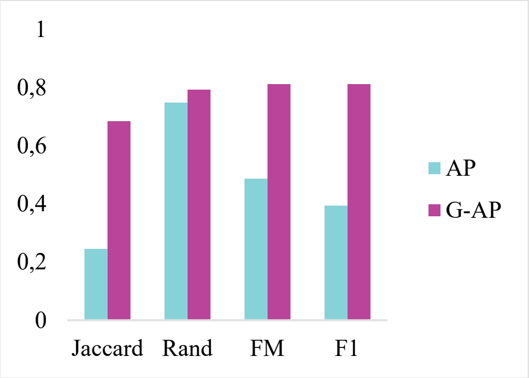

Figure 11 Evaluation index comparison results of iris

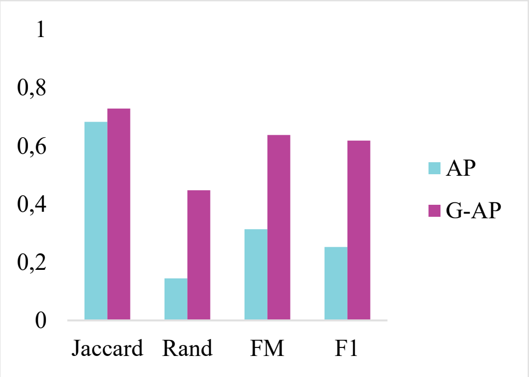

Figure 12 Evaluation index comparison results of wine

According to these evaluation indicator formulas, this paper objectively compare the two algorithms, and obtain the GAP algorithm is better than traditional affinity propagation algorithm with the four evaluation indicators. The validity of the algorithm is evaluated as shown in Figs. 11-14.

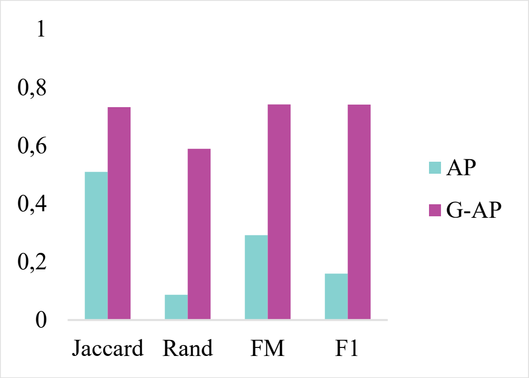

Figure 13 Evaluation index comparison results of flame

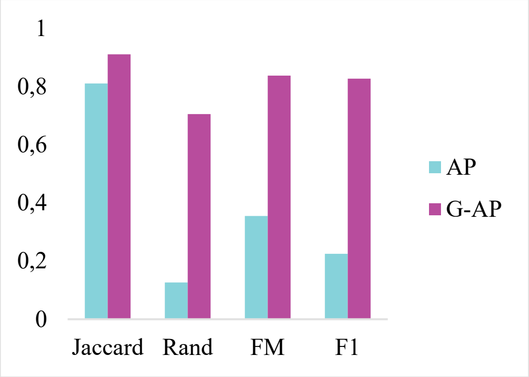

Figure 14 Evaluation index comparison results of aggregation

The effectiveness of the algorithm evaluation results listed in the follow tables.

Table 4 Validity index

Data Set	AP		GAP	
	FM	F1	FM	F1
Iris	0.48	0.39	0.82	0.81
Wine	0.31	0.25	0.64	0.62
Flame	0.29	0.16	0.74	0.74
Aggregation	0.35	0.23	0.84	0.83

Table 5 Validity index

Data Set	AP		GAP	
	Rand	Jaccard	Rand	Jaccard
Iris	0.75	0.25	0.80	0.68
Wine	0.14	0.68	0.45	0.72
Flame	0.09	0.51	0.59	0.71
Aggregation	0.12	0.79	0.76	0.83

4.2 The Analysis of Experimental Results

The simulation results from Tab. 4 , Tab. 5 and Tab. 6 shows that when the GAP algorithm cluster data on basis of the above four data sets, which fully consistent with the data set of real class number [31]. At the same time, through 4 different effectiveness index analysed that the accuracy of clustering algorithm can be found which the 4 data sets have improved significantly [32], it explain that the GAP algorithm can carry on the reasonable data clustering in order to achieve the real clustering requirements, and the clustering effect is better [33, 34].

5 APPLICATION OF THE GAP ALGORITHM IN ANALYSIS OF ECONOMIC DEVELOPMENT IN DIFFERENT REGIONS

5.1 Data Selection

This paper selected six economic indicators that can reflect the economic development of each region: Gross Domestic Product (GDP), The Added Value of The First Industry (AVFI), The Added Value of The Second Industry (AVSI), The Added Value of The Third Industry (AVTI), Consumer Price Index (CPI) and Investment in fixed assets in the whole society (SFAI). In addition, these economic indicators can reflect the development trend of the regions. We can cluster the data, which are the information of the 31 provinces. The paper use these indicators to analyse the development of the region. And obtain the different clusters of these regions, according to the clustering results, we can go on the analysis which related to the development trend of the region, and we can give some advice to the government to develop the region. And the information of the six economic indicators are shown in the follow tables.

In this paper, we used this information to run the GAP algorithm for exploring the application value of the algorithm.

Table 6 Economic indicators

REGION	GDP	AVFI	AVSI	AVTI	CPI	SFAI
Beijing	19500.6	161.8	4352.3	14986.4	103.3	6847.1
Tianjin	14370.2	188.5	7276.7	6905	103.1	9130.3
Hebei	28301.4	3500.4	14762.1	10038.9	103	23194.2
Shanxi	12602.2	773.8	6792.7	5035.8	103.1	11031.9
Neimeng	16832.4	1599.4	9084.2	6148.8	103.2	14215.5
Liaoning	27077.7	2321.6	14269.5	10486.6	102.4	25107.7
Jilin	12981.5	1509.3	6858.2	4613.9	102.9	10133.5
Heilong jiang	14382.9	2516.8	5918.2	5947.9	102.2	12126

Table 7 Economic indicators

REGION	GDP	AVFI	AVSI	AVTI	CPI	SFAI
Hubei	24668.5	3098.2	12171.6	9398.8	102.8	1907.3
Hunan	24501.7	3099.2	11517.4	9885.1	102.5	17846.4
Guang dong	62164	3047.5	29427.5	29689	102.5	22307.9
Guangxi	14278	2343.6	6863	5171.4	102.2	11907.7
Hainan	3146.5	756.5	871.3	1518.7	102.8	2697.4
Chong qing	12656.7	1016.7	6397.9	5242	102.7	10429.6
Guizhou	8006.8	1029.1	3243.7	3734	102.5	7373.6
Yuannan	11720.9	1895.3	4927.8	4897.8	103.1	9968.3
Xizang	807.7	86.8	292.9	427.9	103.6	876
Shanxi	16045.2	1526.1	8911.6	5607.5	103	14867.3
Gansu	6268	879.4	2821	2567.6	103.2	6527.9
Qinghai	2101.1	207.6	1204.3	689.2	103.9	2361.1
Ningxia	2565.1	223	1265	1077.1	103.4	2651.1
Xinjiang	8360.2	1468.3	3766	3126	103.9	7724.5
Sicuan	26260.8	3425.6	13579	9256.1	102.8	20325.2
Shanghai	21602.1	129.3	8027.8	13445.1	102.3	5647.8
Jiangsu	59161.8	3646.1	29094	26421.6	102.3	36373.8
Zhejiang	37568.5	1784.6	18446.7	17337.2	102.3	20777.1
Anhui	19038.9	2348.1	10404	6286.8	102.4	18621.6
Fujian	21759.6	1936.3	11315.3	8508	102.5	15327.4
Jiangxi	14338.5	1636.5	7671.4	5030.6	102.5	12866.1
Shandong	54684.3	4742.6	27422.5	25519.2	102.2	36789.1
Henan	32155.9	4059	17806.4	10290.5	102.9	26220.9

This information are from the 2015 statistical yearbook on the website of the National Bureau of Statistics. This paper use the GAP algorithm to cluster this information, and we can attain the different classes, according to the result, we can compare the result with the reality economic development situation to validate the accuracy of the clustering result and the applicability of the improved algorithm GAP.

5.2 The Analysis of Clustering Result

In order to illustrate the proposed clustering algorithm, it is feasible and practical for regional economic evaluation in our country; this paper respectively used the six different economic indicator values to compare the economic development of the three categories.

Table 8 The analysis of clustering result

Category	The region
First cluster	Beijing, Hebei, Shanghai, Shandong, Jiangsu, Zhejiang, Chongqing, Hunan, Fujian, Guangdong, Tianjin
Second cluster	Anhui, Hubei, Heilongjiang, Jilin, Neimeng, Guangxi, Shanxi, Liaoning, Jiangxi, Sicuan, Henan
Third cluster	Xinjiang, Xizang, Qinghai, Gansu, Ningxia, Shanxi, Hainan, Yunnan, Guizhou

Gross domestic product refers to the market value of all the final products and services produced by all the resident units in a country or region within a certain period. GDP is the core index of national economic accounting, it is also an important indicator of a country's overall economic situation, but not for the measure of a region or city economy, because each city's GDP to the superior or the state are different, so in each city left the wealth is not the same

The GDP refers to the results of production activities of all the permanent units in the region in a certain period. It reflects the economic performance of a region and is the best index to measure the regional economic situation. In addition, we can get that the regions of the cluster1 is higher than cluster2 and cluster3.

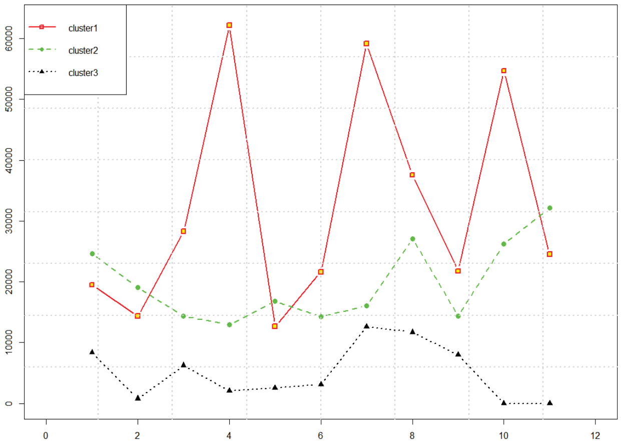

Figure 15 The GDP comparison of the three clusters

The added value of the primary industry, of course, is that the product is directly taken from natural sectors (including farming, forestry, animal husbandry, and

Fisheries) in this liquidation cycle (generally in years) than in the last liquidation cycle, and according to the theory, the AVFI is higher, the degree of urbanization is higher. In addition, the economic development will be better. From Fig. 16, we can get the regions that the degree of urbanization of cluster2 are higher than cluster3 and cluster1, because the regions in cluster2 pay more attention to the agricultural and Forestry Animal Husbandry than the regions in cluster1 and cluster3.

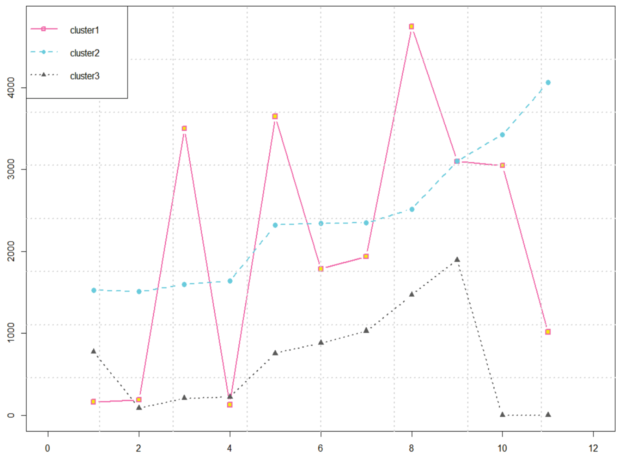

Figure 16 The AVFI comparison of the two clusters

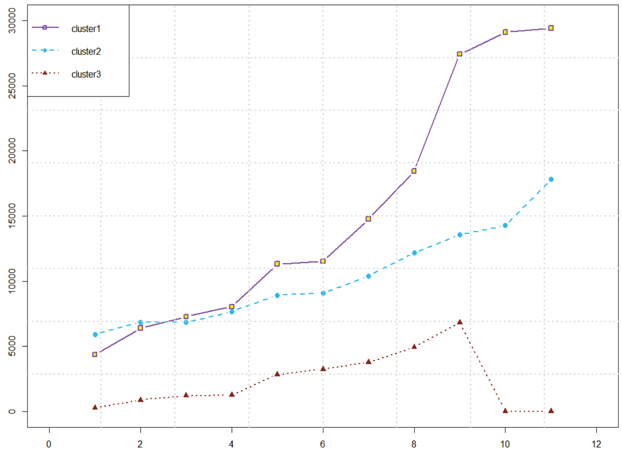

Figure 17 The AVSI comparison of the two clusters

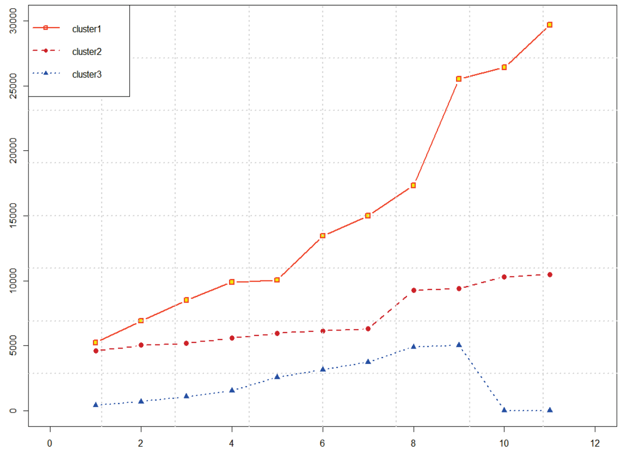

Figure 18 The AVTI comparison of the two clusters

The benefit of secondary industry refers to the results of the secondary industry in monetary terms during the reporting period of production activities; it is the balance of total results of production units of all production activities deduction of the material value of goods and services consumed in the production process or transfer after; it is the newly increased value of the secondary industry in the production process. In addition, according to the theory, we can get that when the AVSI is higher, the industrialization degree of the region is higher. Moreover, the economic development will be better. From Fig. 17, we can get the regions that the degree of industrialization of cluster1 is higher than the others.

Third industries include: transportation, warehousing and postal services, information transmission, computer services and software industry, wholesale and retail, accommodation and catering industry, financial industry, real estate industry, leasing and business services, scientific research, technical services and geological prospecting, water conservancy, environment and public facilities management industry, residents services and other services, education, health, social security and social welfare, culture, sports and entertainment, public administration and social organizations, international organizations and other industries. The third industry has the unity of production and consumption, wide distribution, small dispersion, easy to absorb labour force, to open up the market plays a more and more important role in solving the employment, and its development is good or bad directly affect the national economic development level and quality, reflects part of the effect of whole body and structure optimization can make the overall maximum comprehensive function, conducive to the development and progress of society, is conducive to the improvement of people's living standard.

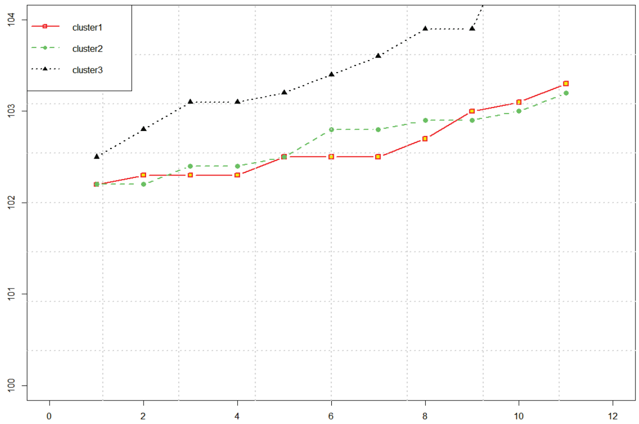

Figure 19 The CPI comparison of the two clusters

CPI is the abbreviation of the consumer price index. The consumer price index is a macroeconomic indicator reflecting the change in the price level of consumer goods and services purchased by the general household. It is a measure of a set of representative goods and services in a specific period that the price level varies with time, and it is used to reflect the households to buy consumer goods and services price level changes. Usually, the CPI is lower, and the economic situation is better. From the Fig. 19, as a whole, the economic situation of the cluster1 is better than cluster2 and cluster3.

The fixed assets investment in the whole society is the workload of building and purchasing the fixed assets in monetary performance. It is a comprehensive index reflecting the scale, speed, proportion and direction of the investment in the fixed assets. In addition, the SFAI is higher, the more possible to stimulate economic growth of the region. From the Fig. 20, as a whole, the ability of the cluster1 is better than cluster2 and cluster3.

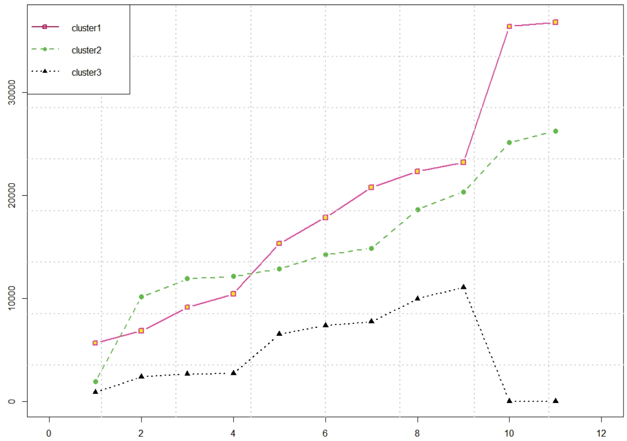

Figure 20 The SFAI comparison of the two clusters

According to the above clustering results, combined with the analysis of the characteristics of financial data, this paper analyses and summarizes the six economic indicators of the 31 provinces. Moreover, the clustering results are in line with the characteristics of regional distribution in China: the eastern coastal areas are developed, and the western regions are underdeveloped.

It can be seen from figures that the regional GDP of the first category is higher than the second. This shows that the economic strength of the first kind of area is relatively strong, the industrial structure is reasonable, while the economic strength of the second areas is relatively weak, and the industrial structure needs to be improved. At the same time our national policy requirements, we should give priority to resource development and infrastructure projects in second areas, second areas to increase poverty alleviation efforts, economic and technical cooperation to strengthen the joint first class area and second area, guiding foreign investment more to second regions.

6 CONCLUSION

In summary, the original Affinity Propagation algorithm existed the drawbacks in calculating similarity [35]. In order to solve these drawbacks, in this paper, we introduced the local density attribute and the idea of universal gravitation, and through the developed similarity calculation to improve the algorithm clustering accuracy and rationality. The proposed algorithm in this paper was applied to the provincial economic data of China, and it fully proved the practicability and application value of the algorithm. Through the example verification, we could find the GAP was not suitable for all type data, and it did not get the perfect clustering result. Thus, in the future, we should improve the AP algorithm in other aspects to make the clustering result of the algorithm be more accurate and application value be greater [36].

Acknowledgments

The research was supported by the national science foundation of China under grant no. 61202306, 61472049, 61402193, 61572225 and 61202309; the national society science foundation of china under grant no. 15bgl090; the foundation of Jilin provincial science & technology department under grant no. 20150312001zg and 20160101259jc; the society science research foundation of Jilin province under grant no. 2014b166; Jilin province education department under grant no. 2014159 and 2016 [78]; the open fund project of logistics industry economy and the intelligent logistics laboratory of Jilin province under grant no.02060311 and 02060312; the fund of Jilin province economic structure strategic adjustment under grant no.2015y040 and 2015y037.

7 REFERENCES

[1] Boulicaut, J. F. & Masson, C. (2017). *Data Mining Query Languages*. Data Mining and Knowledge Discovery Handbook, 655-664.

[2] Buczak, A. L. & Guven, E. (2017). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153-1176. https://doi.org/10.1109/COMST.2015.2494502

[3] Serdah, A. M. & Ashour, W. M. (2016). Clustering Large-Scale Data Based On Modified Affinity Propagation Algorithm. *Journal of Artificial Intelligence & Soft Computing Research, 6*(1), 23-33. https://doi.org/10.1515/jaiscr-2016-0003

[4] Krishnapuram, B., Shah, M., Smola, A. et al. (2016). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM.

[5] Xie, P., Zou, C. W., & Liu, H. (2012). Research on Internet financial model. *New financial review*, 1, 11-22.

[6] Liu, T. T. (2014). *Empirical analysis of the impact of financial position on stock prices in Chinese listed companies*. South-western University of Finance and Economics, 5-6.

[7] Liu, X. J., Chen, M. et al. (2013). The present situation of Chinese stock market. *Technology and market*, 1, 125-125.

[8] Zou, L., Xu, Y., Jiang, Z. et al. (2016). *Functional Connectivity Analysis of Cognitive Reappraisal Using Sparse Spectral Clustering Method*. Advances in Cognitive Neurodynamics (V). Springer Singapore. https://doi.org/10.1007/978-981-10-0207-6_40

[9] Frey, B. J. & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972-976. https://doi.org/10.1126/science.1136800

[10] Fujiwara, Y., Irie, G., & Kitahara, T. (2011). Fast algorithm for affinity propagation. *International Joint Conference on Artificial Intelligence, AAAI Press*, 2238-2243.

[11] Feng, X. & Yu, H., editors. (2011). Semi-supervised affinity propagation clustering based on manifold distance. *Application Research of Computers, 28*(10), 3656-3658.

[12] Wang, L., Ji, Q., & Han, X. (2014). Self-adaptive affinity propagation clustering algorithm based on singular value decomposition. *Journal of Jilin University (Science Edition)*, (04), 753-757.

[13] Zhu, Q., Zhang, H., & Yang, Q. (2015). Semi-supervised Affinity Propagation Clustering Based on Subtractive Clustering for Large-Scale Data Sets. *In: Wang H. et al. (eds) Intelligent Computation in Big Data Era. ICYCSEE 2015.*

Communications in Computer and Information Science, vol 503. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-46248-5_32

[14] Serdah, A. M. & Ashour, W. M. (2016). Clustering Large-Scale Data Based on Modified Affinity Propagation Algorithm. *Journal of Artificial Intelligence & Soft Computing Research, 6*(1), 23-33. https://doi.org/10.1515/jaiscr-2016-0003

[15] Yang, C., Liu, S., Bruzzone, L., Guan, R., & Du, P. (2013). A feature-metric-based affinity propagation technique for feature selection in hyperspectral image classification. *Geoscience and Remote Sensing Letters, IEEE, 10*(5), 1152-6. https://doi.org/10.1109/LGRS.2012.2233711

[16] Wang, L., Han, X., & Ji, Q. (2015). Semi-supervised Affinity Propagation Clustering Algorithm Based on Fireworks Explosion Optimization. *International Conference on Management of E-Commerce and E-Government, IEEE*, 273-279.

[17] Sakellariou, A., Sanoudou, D., & Spyrou, G. (2012). Combining multiple hypothesis testing and affinity propagation clustering leads to accurate, robust and sample size independent classification on gene expression data. *BMC Bioinformatics, 13*(1), 1-19. https://doi.org/10.1186/1471-2105-13-270

[18] Zhang, X., Furtlehner, C., Germainrenaud, C. et al. (2014). Data Stream Clustering With Affinity Propagation. *IEEE Transactions on Knowledge & Data Engineering, 26*(7), 1644-1656. https://doi.org/10.1186/1471-2105-13-270

[19] Lu, Z. & Carreiraperpinan, M. A. (2008). Constrained spectral clustering through affinity propagation. *Computer Vision and Pattern Recognition, CVPR 2008. IEEE Conference on. IEEE Xplore*, 1-8.

[20] Hassanabadi, B., Shea, C., Zhang, L. et al. (2014). Clustering in Vehicular Ad Hoc Networks using Affinity Propagation. *Ad Hoc Networks, 13*(1), 535-548. https://doi.org/10.1016/j.adhoc.2013.10.005

[21] Han, X. H., Quan, L., Xiong, X. Y. et al. (2017). A novel data clustering algorithm based on modified gravitational search algorithm. *Engineering Applications of Artificial Intelligence, 61*, 1-7. https://doi.org/10.1016/j.engappai.2016.11.003

[22] Hatamlou, A., Abdullah, S., & Nezamabadi-Pour, H. (2011). Application of gravitational search algorithm on data clustering. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6954 LNAI, pp. 337-346). https://doi.org/10.1007/978-3-642-24425-4_44

[23] Kumar, Y. & Sahoo, G. (2014). A Review on Gravitational Search Algorithm and its Applications to Data Clustering & Classification. *International Journal of Intelligent Systems & Applications, 6*(6), 79-93. https://doi.org/10.5815/ijisa.2014.06.09

[24] Nikbakht, H. & Mirvaziri, H. (2015). A new algorithm for data clustering based on gravitational search algorithm and genetic operators. *International Symposium on Artificial Intelligence and Signal Processing, IEEE*, 222-227. https://doi.org/10.1109/AISP.2015.7123532

[25] Hatamlou, A., Abdullah, S., & Nezamabadi-pour, H. (2012). A combined approach for clustering based on K -means and gravitational search algorithms. *Swarm and Evolutionary Computation, 6*, 47-52. https://doi.org/10.1016/j.swevo.2012.02.003

[26] Tan, W. S., Hassan, M. Y., Rahman, H. A. et al. (2013). Multi-distributed generation planning using hybrid particle swarm optimisation-gravitational search algorithm including voltage rise issue. *IET Generation Transmission & Distribution, 7*(9), 929-942. https://doi.org/10.1049/iet-gtd.2013.0050

[27] Cheng, K. & Ma, L., editors. (2013). Artificial glowworm swarm optimization algorithm for 0-1 knapsack problem. *Application Research of Computers, 30*(4), 993-994.

[28] Wang, Y. X. & Zhang, Y. J. (2013). Nonnegative Matrix Factorization: A Comprehensive Review. *IEEE Transactions on Knowledge & Data Engineering, 25*(6), 1336-1353. https://doi.org/10.1109/TKDE.2012.51

[29] Li, L. J., Song, K., & Zhao, Y. K. (2011). Modeling of ARA fermentation based on affinity propagation clustering. *CIESC Journal, 62*(8), 2116-2121.

[30] Ma, Y., Peng, M., Xue, W. et al. (2013). A dynamic affinity propagation clustering algorithm for cell outage detection in self-healing networks. *Wireless Communications and NETWORKING Conference, IEEE*, 2266-2270.

[31] Zhang, Z., Wang, B. Q., Yi, P. et al. (2013) Semi-supervised Affinity Propagation Clustering Algorithm Based on Stratified Combination. *Journal of Electronics & Information Technology, 35*(3), 645-651. https://doi.org/10.3724/SP.J.1146.2012.00673

[32] Qi, L., Yu, H., & Min, W. (2015). Active semi-supervised affinity propagation clustering algorithm based on pair-wise constraints. *Intelligent Control and Automation, IEEE,* 2304-2309.

[33] Xu, B., Hu, R., & Guo, P. (2013). Combining affinity propagation with supervised dictionary learning for image classification. *Neural Computing and Applications,* 1-8. https://doi.org/10.1007/s00521-012-0957-7

[34] Graham, E. D., Heidelberg, J. F., Tully, B. J. (2017). BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *Peerj, 5(Part B)*, e3035. https://doi.org/10.7717/peerj.3035

[35] Chen, D. W., Sheng, J. Q., Chen, J. J., et al. (2014). Stability-based preference selection in affinity propagation. *Neural Computing and Applications, 25*(7-8), 1809-1822. https://doi.org/10.1007/s00521-014-1671-4

[36] Zhou, R., Liu, Q., Xu, Z., et al. (2017). Improved fruit fly optimization algorithm based density peak clustering and its applications. *Tehnicki vjesnik, 24*(2), 473-480. https://doi.org/10.17559/TV-20170303013036

Contact information:

Limin WANG
School of Management Science and Information Engineering,
Jilin University of Finance and Economics,
Jilin Big Data Research Center for Business,
Changchun, 130117, China
wlm_new@163.com

Zhiyuan HAO
School of Management Science and Information
Engineering, Jilin University of Finance and Economics,
Jilin Big Data Research Center for Business,
Changchun, 130117, China
1429173931@qq.com

Xuming HAN
Corresponding author
School of Computer Science and Engineering,
Changchun University of Technology,
Changchun, 130117, China
hanxvming@163.com

Ruihong ZHOU
School of Management Science and Information Engineering,
Jilin University of Finance and Economics,
Jilin Big Data Research Center for Business,
Changchun, 130117, China
zrh@jlufe.edu.cn