[1]Yao Wang
[2,*]Fuguo Liu
[3]Guodong Li

Near-neighbor Propagation Clustering Algorithm Based on Cuckoo Search

Journal of Electrical Systems

Abstract: - In this paper, a nearest neighbor propagation clustering algorithm (CSB-AP) based on cuckoo search is proposed to solve the problem of poor parameter setting of the AP algorithm. A population-intelligent optimized cuckoo search algorithm is introduced into the AP algorithm to find the appropriate parameters. Two important parameters in the nearest neighbor propagation algorithm are taken as the location of the bird's nest. The BWP value index is introduced as the bird's nest fitness in the process of intelligent search, and the minimum value is obtained as the final fitness through the reverse mechanism. The final parameters are substituted into the calculation as the best parameters. The CSB-AP algorithm is verified by the UCI data set, and compared with the traditional AP nearest neighbor propagation clustering algorithm, it is found that the CSB-AP algorithm proposed in this paper is better than the traditional AP nearest neighbor propagation clustering algorithm. By comparison, it can be found that the clustering result obtained by the CSB-AP algorithm is closer to the actual result, and it can be known according to the results of the BWP index, contour coefficient, Recall, and F-measure. The improved algorithm can significantly improve the clustering quality and clustering performance.

Keywords: Nearest Neighbor Propagation Clustering Algorithm, Cuckoo Search, Bias Parameter, Convergence Factor, Optimization.

I. INTRODUCTION

Affinity Propagation (AP) algorithm is a new Instance-Based clustering algorithm proposed by Frey and Duec in "Science" in 2007 and is often referred to as the nearest neighbor propagation algorithm [1]. It is characterized by the condition that under the condition of not having to specify the number of clusters in advance, it can quickly and effectively deal with the problems of non-Euclidean space (e.g., not satisfying the symmetry or the triangular inequality) as well as the problem of large-scale sparse matrix computation, and it can obtain the more desirable clustering results. The algorithm is to view all data points as one node, avoiding the clustering results being limited by the choice of initial class representative points. A network (equivalent to a similarity matrix) is then formed between every two nodes in the sample. The messages (Responsibility and Availability) from each edge in the network are transmitted recursively along the node connections until the optimal set of class representative points is found [2].

Neighborhood propagation clustering algorithm is a clustering algorithm that divides data by discovering its intrinsic characteristics. It has many advantages in terms of data processing capability such as high intelligence, faster convergence, and suitable for parallel operations. Because of this, theoretical or applied research in traditional algorithms has become the goal pursued by the majority of researchers. However, there are some shortcomings, and the performance needs to be improved in the processing of massive datasets in the era of big data, and there is still a lot of room for improvement in the parameter settings of bias parameter and damping coefficient. The algorithm's bias parameter is set too large to lead to the final division of the number of clusters being too large, and the damping coefficient is set too large to lead to slow convergence, so the selection of appropriate parameters has a crucial role in the clustering results. On the problem that bias parameters and damping coefficients affect the clustering results, Wang et al. [3] proposed a clustering algorithm with adaptive parameter adjustment. He used the gray wolf pack optimization algorithm to adaptively adjust the parameters and search within a reasonable interval. The values of the parameters are dynamically updated using the binary search algorithm, which can effectively adjust and find the optimal bias parameters, making the number of clusters closer to the real results and improving the quality of clustering. Ma et al. [4] for the influence of bias parameter and damping coefficient on the clustering effect, using the group intelligence algorithm to find the optimal two parameters in the search space, and set the bias parameter and damping coefficient as the brightness and attraction in the firefly algorithm, respectively, which effectively improves the quality of the AP algorithm, but the algorithm utilized in the complex dataset there are still some limitations, and how to effectively carry out the parameter setting there is still room for further research.

[1] School of Mathematics and Data Science, Changji University, Xinjiang, China
[2] School of Mathematics and Data Science, Changji University, Xinjiang, China
[3] School of Mathematics and Computational Science, Guilin University of Electronic Science and Technology Guilin, China
*Corresponding author: Fuguo Liu

II. THE CUCKOO SEARCH ALGORITHM

The cuckoo search algorithm is an emerging heuristic algorithm proposed in 2009 by British scholars Yang et al [5]. It mainly simulates the nest parasitism breeding behavior of cuckoos and the search mechanism of Levy flights. Nest parasitism is a breeding behavior in which some birds lay their eggs in the nests of other birds, which incubate and raise the young on their behalf, and is used by some species of cuckoos to produce offspring. During breeding, cuckoos do not build nests, but instead search for hosts with similar incubation and brood stages, similar chick diets, and similar egg shapes and colors. After finding a nest with a suitable host, they will quickly lay their eggs while the host is away from the nest and allow the host bird to incubate the young instead of themselves. Typically, cuckoos lay only one egg in a single parasitic nest, and they will remove the host bird's eggs, or all of them, from the nest before doing so. Levy flight is a kind of foraging random wandering, the wandering step length satisfies a heavy-tailed stable distribution, the short-distance exploration is interspersed with occasional long-distance walks, and it is also one of the most optimal foraging searching ways to find the optimal nests to incubate their own birds' eggs, and use this way to achieve an efficient optimality seeking mode, and use Levy searching way in intelligent algorithms to be able to overcome the easily fall into the shortcomings of the local optimal solution, the algorithm has few parameters, simple operation, easy to implement, stochastic search path ability and optimality seeking ability, which is highly concerned by scholars [6].

The pseudo-code for the cuckoo algorithm is shown as follows.

Step 1: Set the fitness $F(s)$, $s = (s_1, s_2, s_3, \dots, s_d)^T$, initialize the number and location of bird nests.

Step 2: Calculate the fitness $F(s)$ and record when the current optimal solution is encountered.

Step 3: The position of the dataset (bird's nest) is updated by the criteria of the Levy flight mechanism.

Step 4: Recalculate the fitness $F(s)$ of the dataset location by comparing it with the last recorded location information. Determine the current locally optimal solution;

Step 5: A uniform distribution $p \in [0,1]$ is used and its value is used as a probability of being found to be a bird's egg from a non-nest owner, which is compared with $p\alpha$. If $p > p\alpha$, then the location of the dataset is randomly changed.

Step 6: If the iteration condition for exiting the algorithm is contented, the optimal solution; if not, return to step 2.

III. PARAMETRIC ANALYSIS OF THE NEAREST NEIGHBOR PROPAGATION ALGORITHM

A. Bias Parameter

The value of the bias coefficient is the element of the main diagonal in the similarity matrix $s(i,k)$, and the algorithm assumes that each data point may be a class representative point, then the probability of being the center of the clustering is the same, and sets the same value of the bias parameter which is p [7].

B. Damping Factor

In the AP algorithm, λ is generally a fixed value, the larger λ is, the slower the convergence rate is, and the value ofλ affects the convergence rate of the algorithm and the stability of the iteration [8].

C. Clustering Effect of Different Parameters

In response to the uncertainty in the parameters of the AP algorithm, this paper will use two data from the UCI dataset for the experiments. Different values of p and different values of λ will be selected. The bias coefficient p is selected from 1 times pmed to 5 times pmed respectively. setting the value of the damping coefficient p is generally set as the principle above the median or the minimum $s(i,k)$, and the structure of different datasets is also different. Therefore, in this paper, we set different damping coefficients for different data, and generally selectλ between the range of 0.5 and 0.8.

For the wind data in Table 1, the damping coefficients λ are set to 0.5, 0.6, and 0.7, and the bias coefficients are chosen to be the values of p for 3 times pmed, 4 times pmed, 5 times pmed, and 6 times pmed, from the above four data comparisons, setting different parameters will have different clustering effects. Regardless of the bias argument, the clustering effect of the algorithm is not good when the damping coefficients are 0.5 and 0.6; the clustering effect is better when the bias parameter is 0.7, and the clustering effect is optimal when the bias parameter is -2400 from the view of the clustering evaluation index indicator. Numerous experiments have shown that the clustering effect is best when the bias parameter and damping coefficient are set at -2400, 0.7. To summarize: different parameters will make the algorithm have different effects, and the correct parameter settings in the AP algorithm will affect the clustering effect.

Table 1: Clustering Effect of Wind Dataset under Different Parameters

Bias parameter	Damping factor	Number of clusters	Homogeneity	Completeness	V-measure	Adjusted Rand Index	Adjusted Mutual Information	Silhouette Coefficient
-1200	0.5	1	0	1	0	0	0	0
-1200	0.6	4	0.423	0.334	0.373	0.347	0.325	0.131
-1200	0.7	3	0.366	0.363	0.364	0.369	0.356	0.244
-1600	0.5	177	1	0.21	0.347	0	0	0
-1600	0.6	4	0.44	0.346	0.387	0.334	0.337	0.114
-1600	0.7	3	0.378	0.373	0.376	0.377	0.367	0.247
-2000	0.5	133	0.99	0.245	0.393	0.116	0.073	-0.017
-2000	0.6	3	0.381	0.378	0.38	0.376	0.371	0.258
-2000	0.7	3	0.359	0.354	0.357	0.371	0.348	0.214
-2400	0.5	120	0.84	0.224	0.354	0.115	0.053	-0.322
-2400	0.6	3	0.394	0.39	0.392	0.385	0.383	0.262
-2400	0.7	3	0.381	0.378	0.38	0.376	0.371	0.258

Table 2: Clustering Effect of Seed Dataset under Different Parameters

Bias parameter	Damping factor	Number of clusters	Homogeneity	Completeness	V-measure	Adjusted Rand Index	Adjusted Mutual Information	Silhouette Coefficient
-600	0.5	5	0.784	0.547	0.644	0.573	0.541	0.36
-600	0.6	4	0.763	0.619	0.684	0.678	0.615	0.56
-600	0.7	5	0.76	0.532	0.626	0.569	0.527	0.371
-1200	0.5	65	0.881	0.319	0.469	0.382	0.232	-0.268
-1200	0.6	3	0.696	0.697	0.697	0.738	0.694	0.576
-1200	0.7	4	0.694	0.555	0.617	0.566	0.55	0.283
-1800	0.5	60	0.781	0.335	0.469	0.452	0.243	0.141
-1800	0.6	3	0.707	0.712	0.709	0.710	0.704	0.547
-1800	0.7	3	0.699	0.700	0.699	0.739	0.696	0.577
-2400	0.5	149	0.948	0.232	0.373	0.089	0.07	-0.211
-2400	0.6	3	0.730	0.732	0.731	0.763	0.728	0.620
-2400	0.7	3	0.708	0.708	0.708	0.750	0.705	0.599

It is found that the lower the damping coefficient, further improvement the clustering effect through the test of six clustering evaluation indexes. Through the above analysis, it can be concluded that the correct setting of the two parameters of bias coefficient p and damping coefficient λ of the AP algorithm will be crucial to the results of the clustering algorithm. As for the two datasets, Iris and Seed, it is found through Tables 1 and 2 that when the damping coefficient λ is a fixed value, the bias coefficient p changes monotonically in the process of the evaluation index. Similarly, the bias coefficient p exhibits the same variation when it is a fixed value, with similar conclusions. The findings from a large number of experiments indicate that different values of p affect the value of F when the value of λ is a fixed value. different values of λ affect the value of λ when the value of p is a fixed value. In summary, setting different values of p or different values of λ affects the clustering effect, and the effect will be evaluated by using evaluation indexes to judge whether the clustering effect is good or bad.

IV. Optimization of Neighborhood Propagation Clustering Algorithm

Let there exist different clustering classes $c = \{c_1, c_2, c_{3,\ldots,} c_k\}$ in the sample space with sample number $N = \{x_1, x_2, x_3, \ldots, x_n\}$ and each data point corresponds to, and only corresponds to, one clustering class, then the error function of clustering will be defined as Eq. 1 [9].

$$J(C) = \sum_{i=1}^{n} d^2\left(x_i, x_{c(i)}\right) \quad (1)$$

The goal of the this algorithm is to find the best set of class representative points that minimizes the error function. That is, $C = \text{argmin}[J(C)]$.

The similarity $s(i,k)$ between any two sample points x_i and x_j is measured by the negative Euclidean distance, and its value is stored in an $N \times N$ similarity matrix with the mathematical expression Eq. 2.

$$s(i,k) = -d^2(x_i, x_k) = -|x_i - x_k|^2, i \neq k \quad (2)$$

The elements on the diagonal of the matrix are the bias parameter p. The nearest neighbor propagation clustering algorithm initially assumed the same possibility as representing the class, i.e., all the $s(k,k)$ are set to be of the same value p. Generally the value of p is the median of all the values of the similarity matrix, and the magnitude of the value of p influences the number of clusters, increasing the value of p will increase the number of classes. Decreasing the p value decreases the number of classes and is calculated as Eq. 3.

$$P = median(s(\cdot)) \quad (3)$$

The core of the AP algorithm is the information transfer between the sample points to each other, the AP algorithm has two kinds of information, they are attraction (Responsibility) and attribution (Availability) to establish the process as shown in Figure 1 and Figure 2 [10].

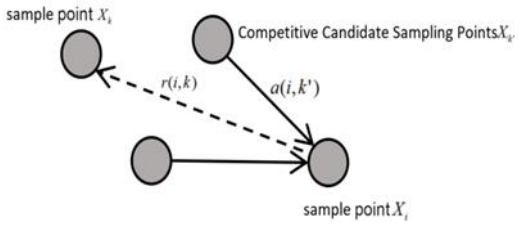

Figure 1: Attractiveness Establishment

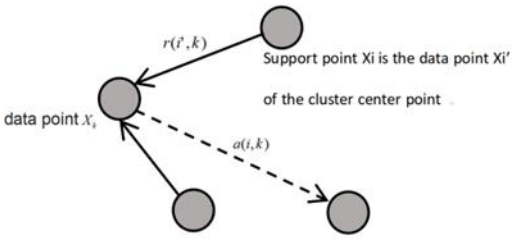

Figure 2: Establishment of the Degree of Belonging

The initial values of both attraction and attribution of the algorithm are zero, indicating that there is no clustering relationship between the samples at the beginning and will be updated as follows. To find a suitable clustering center x_k, evidence $r(i,k)$ and $a(i,k)$ are continuously collected from the data samples. $r(i,k)$ denotes the amount of information that sample point x_i passes on to sample point x_k, i.e., the degree of support that x_i chooses x_k as its clustering center (attraction degree) [11]. $a(i,k)$ denotes the amount of information that sample point x_k sends to sample point x_i, i.e., the degree of fitness that x_i chooses x_k as the fitness of its clustering center (degree of attribution). The process of finding the optimal clustering center by continuously passing updates between samples through the two messages, attribution degree, and attraction degree, is mathematically expressed as follows.

$$r(i,k) \leftarrow s(i,k) - \max_{k's.t.k' \neq k}\{a(i,k') + s(i,k')\} \tag{4}$$

$$a(i,k) \leftarrow \begin{cases} min\{0, r(k,k) + \sum_{k's.t.i' \notin \{i.k\}} max\{0, r(i,k')\}\}, & i \neq k \\ \sum_{i's.t.i' \neq .k} max\{0, r(i',k)\}\ , & i = k \end{cases} \tag{5}$$

$$\arg\ \max_{k}(a(i.k) + r(i,k)) \tag{6}$$

To avoid the oscillation of the nearest neighbor propagation algorithm, a damping factor $\lambda \in [0,1]$ is introduced during the information updating process, and the updated $r(i,k)$ and $a(i,k)$ are obtained by weighted summation of the current values of $r(i,k)$ and $a(i,k)$ with the results of the previous iteration, respectively. In this paper, $\lambda = 0.5$ is chosen and the update formula is as follows.

$$r^{t+1}(i,k) \leftarrow (1-\lambda)r^{t+1}(i,k) + \lambda r^{t}(i,k) \tag{7}$$

$$a^{t+1}(i,k) \leftarrow (1-\lambda)a^{t+1}(i,k) + \lambda a^{t}(i,k) \tag{8}$$

The current number of iterations, when class representation points remain unchanged in several consecutive iteration steps or reaches the maximum number of iterations, the iteration will be terminated and the algorithm will end.

Aiming at the problem that the AP algorithm cannot accurately locate the value of the bias coefficient (p) and damping coefficient (λ), this paper introduces the cuckoo algorithm to improve the searching ability and convergence; the bias coefficient and damping factor are used as the bird's nest, and the CSB-AP algorithm is run to search for the optimal value automatically, to improve the clustering accuracy of the AP algorithm.

The cuckoo search algorithm can be described using the following three ideal rules: (1) Each cuckoo is set to lay only one egg at a time, and the nest into which the egg is placed is chosen at random. (2) The optimal nest that can best incubate the egg is reserved. (3) The number of host nests available for egg placement that the cuckoo can select is fixed, and the probability that a host discovers that an egg placed by a cuckoo is non-parental is $p\alpha \in (0,1)$.

In the CSB-AP algorithm proposed in this paper, firstly, N initial solutions are randomly generated, i.e., the location of the bird's nest, and the fitness of the bird's nest are calculated. By updating the new solutions generated iteratively, the fitness of the new solution is calculated and compared with the fitness of the original solution, and the value with higher fitness is selected as the solution. Then compare the size of the fitness BWP value of the candidate solution with the value of the discovery probability, if the fitness is smaller than the discovery probability, the solution is discarded; when the number of discarded solutions is equal to the number of newly generated candidate solutions, the fitness is utilized for screening, and the solution with the highest fitness value is selected. If the number of iterations meets the termination condition, the CSB-AP algorithm completes the clustering and outputs the results of the run; otherwise, the algorithm continues to run.

The formula for a cuckoo to find a nest and update the nest location is shown below.

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \cdot Levy(\lambda) \tag{9}$$

Where $\alpha > 0$ is the step size, take $\alpha = 1$. Levy( ) is the randomized search path, and the randomized step size formula is $\text{Levy}(\mu) = t^{-\lambda}, 1 < \lambda < 3$.

CSB-AP algorithm flow:

Inputs: similarity matrix $s(i,k)$, the maximum number of iterations T, number of nests N.

Output: number of clusters k, bias coefficient (p), damping coefficient (λ).

Step 1: Randomly select N bird nests as the initial solution, i.e., randomly select the values of bias parameter and convergence factor for each group, but each nest has only and only one bird egg.

Step 2: Build the similarity matrix $s(i,k)$, initialize $r(i,k) = 0$ and $a(i,k) = 0$.

Step 3: Update the two parameters of the AP algorithm, i.e., bias coefficient (p p) and damping coefficient (λ), according to equation $x_i^{(t+1)}$.

Step 4: Run the AP algorithm and use the BWP value as the bird's nest adaptation.

Step 5: Record the value with the lowest fitness, i.e., the location of the bird's nest, whose value is the optimal parameter value (p and λ);

Step 6: If the termination condition of the algorithm is reached, stop; otherwise revert to Step 3.

Pseudo-code for the algorithmic framework, as shown in Figure 3.

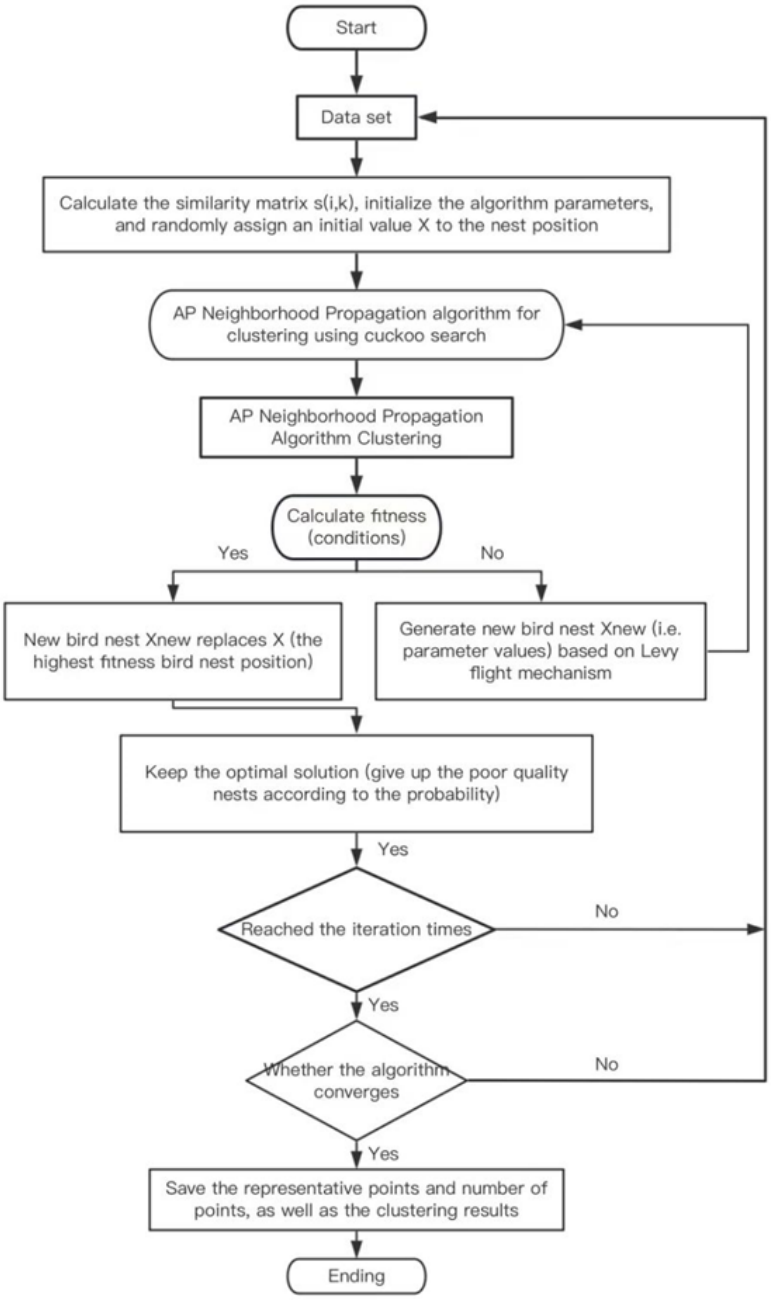

Figure 3: Flowchart of CSB-AP Algorithm

V. COMPARISON AND ANALYSIS OF RESULTS

Experimental environment: this algorithm experiment build platform uses a Core i3720M processor (1.80 GHz), 4.00 GB RAM, using Python as a Programming Language.

A. Experimental Data

To validate the feasibility and effectiveness of the CSB-AP algorithm proposed in this paper, the text will be validated by selecting a synthetic UCI dataset specifically designed to test the performance of clustering algorithms, which is a commonly used standard test dataset with explicit classifications, so the quality of the clusters can be directly observed. In this paper, three datasets with different data sizes are selected for testing. Table 3 shows the basic characterization of the relevant experimental datasets.

Table 3: UCI Experimental Dataset

Datasets	Name	Sample size	Dimension	Class number
Data1	Seed	210	7	3
Data2	Iris	150	4	3
Data3	haberman	306	3	2

B. Experimental Results and Analysis

To verify the effectiveness of the algorithm, the CSB-AP algorithm proposed in the text is compared with the traditional AP algorithm, and the experimental results are shown in the following table.

Table 4: Comparison of the Number of Clusters in Different Clustering Algorithms

Datasets	Number of real classes	AP	CSB-AP
Seed	3	15	2
Iris	3	7	3
haberman	2	26	2

From Table 4, the following conclusion can be drawn: in the three selected datasets, the number of clusters derived from running the CSB-AP algorithm differs less from the actual number of clusters, while the final number of traditional AP algorithms differs more from the standard number. In comparison, the CSB-AP clustering algorithm proposed in this paper has a better performance, indicating that the cuckoo algorithm plays an important role in parameter optimization and better guides the clustering process.

Table 5: Comparison of Evaluation Metrics for Clustering Results for the SEED Dataset

DATA1	AP	SCB-AP	Growth value	Growth rate
BWP indicator	0.029968	0.310188	0.28022	90%
Contour coefficient	0.038861	0.350733	0.31187	88%
recall	0.136365	0.667849	0.53148	80%
F-measure	0.197647	0.689781	0.49213	71%

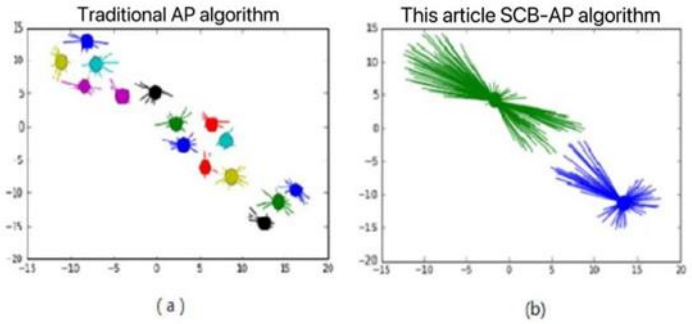

Figure 4: Comparison of Algorithms for the SEED Dataset

According to the results of the comparison of the evaluation indexes of the clustering results of the SEED dataset in Table 5, it is found that the initial fitness of -0.036524 is obtained after randomly setting the parameters using the CSB-AP algorithm proposed in this paper. Through the Levy flight mechanism, 20 points are flown each time for optimization iteration, and the first round of optimal fitness is obtained as -0.049783. The second round of optimal fitness is -0.049969. The third round optimal fitness is -0.053141. The fourth round optimal fitness is -0.055326. The fifth round optimal fitness is -0.053141. The sixth round optimal fitness is -0.055326. The seventh round optimal fitness is -0.055399. The eighth round optimal fitness is -0.059619. The ninth round optimal adaptation is -0.066677. The tenth round optimal fitness is -0.113793. The eleventh round optimal fitness is -0.142124. The twelfth round of the optimum has an adaptation of -0.309191. As the iteration increases, the fitness

is finally minimized with a value of -0.310188 The iteration is stopped to obtain the optimal bird nest location and the final clustering result is obtained based on the selected parameters.

In the comparison of the clustering evaluation metrics based on the two algorithms, all five metrics have been improved significantly. Among them, the BWP index, contour coefficient, and recall three indexes are especially obvious, which are nearly improved by 10 times. As shown in Figure 4, from the clustering comparison chart of the two algorithms, it is found that according to the Levy flight mechanism, the SEED dataset changes from 15 classes to two classes, which is closer to the standard class number. It is not difficult to find that this up-to-date algorithm proposed is superior to the old AP algorithm. It can be seen that the cuckoo search algorithm can accurately search out the optimal parameter values (damping coefficient and bias parameter), so that the final clustering result achieves the effect of a tighter intracluster and more distant interclass.

Table 6: Comparison of Evaluation Metrics for Clustering Results for the Iris Dataset

DATA2	AP	SCB-AP	Growth value	Growth rate
BWP indicator	0.086364	0.279259	0.192895	69%
Contour coefficient	0.103200	0.298191	0.194991	65%
recall	0.268418	0.670034	0.401616	60%
F-measure	0.317577	0.633684	0.316107	50%

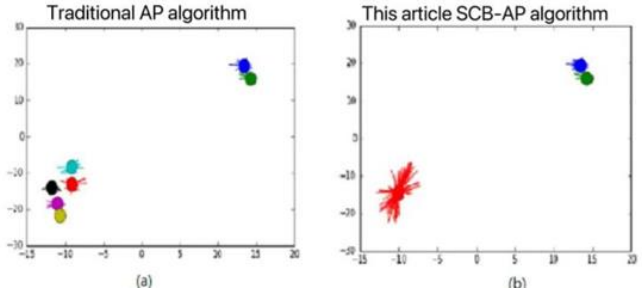

Figure 5: Comparison of Algorithms for the Iris Dataset

According to Table 6 Iris dataset clustering results evaluation metrics comparison results are found in the initial fitness obtained after randomly setting the parameter values as -0.062756. according to the Levy flight mechanism, the first round of the optimal fitness is calculated to be -0.128501. each time the flight is iterated with 20 points, the next round of fitness is calculated and compared with the existing solution. According to this mechanism keep cycling to get the third round with an optimal fitness of -0.129261 and the final fitness is -0.279259. The number of iterations has increased, the fitness drops to the lowest value, and the iterations are stopped. The optimal bird's nest location, i.e., the appropriate parameter setting, is found and substituted into the algorithm to obtain the clustering result.

According to the clustering result evaluation indicators of the Iris data set in Figure 5, it is found that the five categories of indicators have greatly improved, indicating that the clustering effect has been improved. It can be found from the clustering diagram that the number of clusters has changed from seven categories to three categories, which is consistent with the standard number, illustrate that the clustering effect of the new algorithm get a jump on the traditional algorithm.

Table 7: Comparison of Clustering Result Evaluation Indicators of Haberman Data Set

DATA3	AP	SCB-AP	Growth value	Growth rate
BWP indicator	0.018651	0.312341	0.29369	94%
Contour coefficient	0.023667	0.369994	0.346327	93%
recall	0.091270	0.555748	0.464478	83%
F-measure	0.162362	0.633347	0.470985	74%

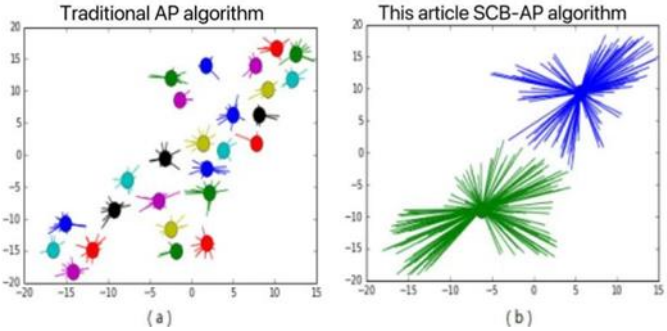

Figure 6: Algorithm Comparison Chart of Haberman Data Set

According to the comparison results of the clustering result evaluation indicators of the haberman data set in Table 7, it is found that similar to the above mechanism, the initial fitness of this data is -0.015590 through experiments. Iterating according to the flight mechanism, the optimal fitness in the first round is calculated to be -0.025661. The optimal fitness in the second round is -0.028107. The optimal fitness in the third round is -0.028663. The optimal fitness in the fourth round is -0.029572. The optimal fitness in the fifth round is -0.029889. The optimal fitness in the sixth round is -0.031191. The optimal fitness in the seventh round is -0.090974. The optimal fitness in the eighth round is -0.181013. The optimal fitness in the ninth round is -0.312341. The optimal fitness in the tenth round is -0.312341. The final fitness is -0.312341. Every increase in the number of iterations, the fitness drops to the lowest value and the iteration ends. According to Figure 6, it is shown that the clustering performance of the new algorithm is better than that of the traditional algorithm.

To summarize, six standard datasets in the UCI database were used for testing. The evaluation index of clustering results shows that the proposed CSB-AP algorithm outperforms the traditional AP algorithm in clustering accuracy, cluster tightness, and computing efficiency. Levy mechanism and cuckoo search algorithm were employed to accurately calculate the nest locations and obtain the values of damping coefficient and bias parameter in AP algorithm, avoiding the uncertainty and complexity introduced by manual setting. Therefore, the proposed CSB-AP algorithm is more effective and efficient in clustering than traditional AP algorithms.

VI. CONCLUSION

The nearest neighbor communication clustering algorithm in coping with the complexity of large data sets in the era of big data, an improved algorithm based on the group intelligence of the cuckoo search method is introduced. The algorithm combines the nearest neighbor communication clustering algorithm with the idea of brood parasitism of the cuckoo and locates the two parameters of the AP algorithm by finding the optimal nest, so that the new algorithm can quickly and accurately locate the optimal bias parameter and damping coefficient, improve the algorithm's ability to search for the global optimum solution, and obtain a better objective value, thereby enhancing the clustering effect. The algorithm uses Euclidean distance as a similarity measure and is more effective in processing data structures that are spherical or approximately spherical. The simulation results show that the proposed optimization algorithm has good performance.

REFERENCES

[1] Frey B J, Dueck D. Clustering by passing messages between data points. science, 2007, 315(5814): 972-976.

[2] Zhou S ,Xu Z .Automatic grayscale image segmentation based on Affinity Propagation clustering.Pattern Analysis and Applications, 2020, 23(1): 331-348.

[3] WANG M L ,ZHOU Y ,HAN M X , et al.Constraint Rules and Matching Micro-clusters Based Affinity Propagation Clustering Algorithm.Studies in Informatics and Control,2020,29(3):353-362.

[4] Zhihong O ,Lei X ,Feng D, et al.Automatic Aggregation Enhanced Affinity Propagation Clustering Based on Mutually Exclusive Exemplar Processing.Electronic Countermeasure Institute, National University of Defense Technology ,Hefei, 230037 ,China,2023,77(1):983-1008.

[5] Bi X ,Guo B ,Shi L , et al.A New Affinity Propagation Clustering Algorithm for V2V-Supported VANETs.IEEE Access,2020,871405-71421.

[6] Bingqi L ,Jianbo H ,Yingyang W , et al.An expert weighting method based on affinity propagation clustering algorithm.Journal of Physics: Conference Series,2019,1324012006-012006.

[7] Biqi L ,Fuxiang Z ,Xi L , et al.Power load identification based on Long-and-Short-Term Memory network and Affinity Propagation clustering algorithm.Energy Reports,2022,8(S4):1137-1144.

[8] Yulong D ,Yang L ,Wuxu P , et al. 3D pseudo-lithologic modeling via iterative weighted k-means++ algorithm from Tengger Desert cover area, China.Frontiers in Earth Science,2023,11

[9] Seju P ,Shin H J ,Cheol M , et al.Dynamic Small-Cell Clustering Using Affinity Propagation Algorithm in Asynchronous 5G NR OFDM Systems.IEEE COMMUNICATIONS LETTERS,2021,25(11):3629-3633.

[10] Seju P ,HanShin J ,Cheol M , et al.RRH Clustering Using Affinity Propagation Algorithm with Adaptive Thresholding and Greedy Merging in Cloud Radio Access Network.Sensors,2021,21(2):480-480.

[11] Yajun Z ,Jie D ,Kangkang Z , et al.Location and Expansion of Electric Bus Charging Stations Based on Gridded Affinity Propagation Clustering and a Sequential Expansion Rule.Sustainability,2021,13(16):8957-8957