Data Clustering Method Based on Bat Algorithm and Parameters Optimization

Ling-Feng Zhu, and Jie-Sheng Wang *

Abstract—**The process of dividing a collection of physical or abstract objects into multiple classes of similar objects is called clustering. Bat algorithm is a new swarm intelligent optimization algorithm, which is proposed to simulate the foraging behavior of bats and their echolocation ability. The bat algorithm is adopted to solve the data clustering problem and the efficient parameters setting on the bat algorithm is carried out the related research on the pulse rate and pulse loudness. The simulation experiment results show that the reasonable setting parameters can improve the performance of the bat algorithm on the clustering problem.**

Index Terms—**clustering, bat algorithm, parameter optimization**

I. Introduction

CLUSTERING is a process of dividing the sample data into clusters according to a certain rule, so that the samples in the same cluster have very high similarity. However, the differences between samples in different clusters are very high. At present, the clustering method is largely unsupervised, and its clustering result is likely to be very different from the actual situation. Clustering analysis as a method of unsupervised learning in recent years has been widely used in machine learning, data mining, artificial intelligence, image processing and other fields, which gradually becomes a hot research spot in this field. Relative to the unsupervised clustering analysis, a semi-supervised clustering method can more effectively improve the clustering performance by combining a small amount of samples and a large number of unmarked tagged samples. The traditional clustering analysis methods include the partitioning method, the hierarchical method, the density-based method, the grid-based method and the model-based method [1-5].

Many of the existing semi-supervised clustering algorithms are often susceptible to the initial points and are prone to the local optima. With the development of modern computer technology and the continuous improvement of the intelligent optimization algorithm, many heuristic algorithms are put forward and are applied to the clustering problems, such as genetic algorithm, ant colony algorithm, particle swarm optimization algorithm, Tabu search algorithm, etc. [4-7]. Bat algorithm (BA) is a heuristic search algorithm proposed by professor Yang in 2010 based on the swarm intelligence thought, which is an effective method to search the global optimal solution [6-9]. BA has the advantages of simplicity, less parameters, strong robustness and easy implementation. It has been widely used in many fields, such as the constrained optimization problem, the fuel arrangement optimization of reactor core, the optimal sizing of battery energy storage, the muti-variable PID controller tuning, the community detection in complex networks, the reserve constrained dynamic economic dispatch problem, and the muti-area load frequency control [10-16]. This paper uses the bat algorithm to solve the data clustering problem, mainly discussing the parameter optimization initialization problem of BA, and verifying the importance of reasonable parameter setting of BA through simulation experiments.

Manuscript received September 24, 2018; revised December 30, 2018. This work was supported by the the Basic Scientific Research Project of Institution of Higher Learning of Liaoning Province (Grant No. 2017FWDF10), and the Project by Liaoning Provincial Natural Science Foundation of China (Grant No. 20180550700).

Ling-Feng Zhu is a postgraduate student in the School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, PR China (e-mail: 1360967798@qq.com).

Jie-Sheng Wang is with the School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, 114051, PR China; National Financial Security and System Equipment Engineering Research Center, University of Science and Technology Liaoning. (Corresponding author, phone: 86-0412-2538355; fax: 86-0412-2538244; e-mail: wang_jiesheng@126.com).

II. Bat Algorithm for Solving Clustering Problem

A. Basic Principle of Bat Algorithm

Bat algorithm (BA) is a new swarm intelligent optimization algorithm to simulate the foraging behavior of bats, whose principle is to use bats' advanced echolocation ability [9]. Echolocation is a sonar. Bats (mainly small bat) produce the loud and short pulse sound. Until the sound arrives at an object, in a relatively short time the echo will return to their ears. So bats can receive and detect the location of the prey in this way. In addition, this directional mechanism enables bats to distinguish between a barrier and prey, allowing them to hunt even in complete darkness. This algorithm adopts the frequency tuning technology to enhance the species diversity in solution. The automatic scaling technology is adopted to dynamically change the transmission rate of the pulse and pulse loudness to balance the searching global optimization and local optimization of the algorithm.

In order to simulate the behavior of bat foraging, the biological mechanism of bat algorithm is assumed idealized as follows:

(1) All bats use echolocation to detect distance, and it identifies food and the surrounding obstacles in a way that we can't understand.

(2) The bats use variables, such as wave λ, loudness A_0, and fixed frequency $f_{\min}$, to search for prey with free flight

manner at location X_i with speed V_i . The bats can dynamically adjust the pulse wavelength (or frequency) based on the distance between the prey and itself, and adjust the frequency $r \in (0,1)$ of the transmitted pulse when it approaches the prey.

(3) Assume that the loudness changes from the maximum A_0 (positive) to the minimum $A_{\min}$ (constant).

On the basis of these three idealized assumptions, the bat algorithm randomly generates a set of solutions, and then searches the optimal solution in the process of searching the optimal solution to strengthen the local search. The random flight near the optimal solution produces the local solutions and finally finds the global optimal solution. The bat's foraging space is d -dimension. At $t-1$, the position and flight speed of i th bat are X_i^{t-1} and V_i^{t-1}, and the current global optimal position is X*. The position and flight speed of i th bat at t moment are updated by the following equations.

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta \tag{1}$$

$$V_i^t = V_i^{t-1} + (X_i^{t-1} - X_*)f_i \tag{2}$$

$$X_i^t = X_i^{t-1} + V_i^t \tag{3}$$

where, $f_{\min}$ and $f_{\max}$ are the minimum and maximum frequencies of sound waves emitted by bats respectively, and β is a uniform random number in [0, 1]. When setting up the initial parameters, the frequency of the transmitted sound waves of each bat obeys the uniform random distribution of in $[f_{\min}, f_{\max}]$, that is to say firstly the respective frequency is set according to Eq. (1), then the speed and position are updated according to Eq. (2)-(3).

For the local searching, every bat has an optimal solution. Its new solution is a random walk close to the optimal solution, which is generated based on the following equation.

$$X_{new} = X_{old} + \varepsilon A^t \tag{4}$$

where, ε is a random number in [-1,1], X_{old} is a solution randomly selected from the current optimal solution, and A^t is the average loudness of all bats when the iteration number is t.

The updating rules of the bat-shot's loudness A_i and rate r_i are described as follows. Assuming that the bat only finds prey, it will reduce the response of its emitted pulse and increase the rate of its emitted pulse. In the bat algorithm, the loudness A_i and rate r_i of the launched pulse are adjusted by:

$$A_i^{t+1} = \alpha A_i^t \tag{5}$$

$$r_i^{t+1} = r_i^t[1 + \exp yt] \tag{6}$$

where, r_i^0 is the initial randomly selected rate, A_i^0 is the initial randomly selected loudness, α and y are constants ($0 < \alpha < 1, y > 0$).

B. Bat Algorithm for Clustering Problem

The clustering analysis is generally considered as a per-processing knowledge discovery process. The whole process of knowledge discovery depends on the quality of clustering. In the process of data mining, clustering also requires effective accuracy. The traditional clustering algorithm is essentially a local search algorithm, and its essence uses an iterative optimization method to find the optimal value. As a result, there are two drawbacks: one is not easy to handle the big data, and the other is that it is easy to produce the local minim.

For clustering problems, the accuracy and optimization of the algorithm have certain limitations on single clustering result. Therefore, many researches combine various algorithms to carry out the cluster analysis. Bat algorithm is a new type of swarm intelligence algorithm, which can adjust the loudness A_i and rate r_i of the emitted pulses to carry out the optimized clustering algorithm, which can prevent the cluster itself falling into the local minimum, and at the same time save the clustering time. The experimental results show that the optimization effect of bat algorithm on data clustering has certain advantages. In this experiment, two sets of data sets are adopted. The data is composed of 200 data matrices with two types of data. One type of data is the training data and the other is the verified data.

C. Pseudo-code of Bat Algorithm

The pseudo-code of bat algorithm is described as follows [16]：.

Objective function $f(X), X = (x_1, ..., x_d)^T$
Initialize the bat population $X_i (i = 1,2...n) and \ V_i$
Define pulse frequency f_i at X_i
Initialize the pulse rates r_i and the loudness A_i
while (t <Max number of iterations)
Generate new solutions by adjusting frequency, and update the velocities and locations/solutions by using Eq. (3) -(4)
if ($rand > r_i$)
Select a solution among the best solutions
Generate a local solution around the selected best solution
End if
Generate a new solution by flying randomly
if ($rand < A_i$ & $f(X_i) < f(X_*)$)
Accept the new solutions
Increase γ_i and reduce A_i use Eq. (5) and (6)
End if
Rank the bats and find the current best X_*
End while

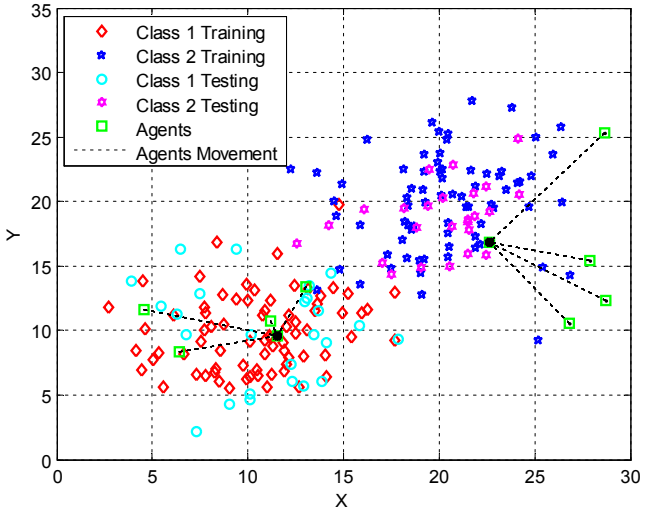
III. Simulation Experiments and Result Analysis

It can be seen from the implementation process of the above described BA, two parameters in the bat algorithm (the attenuation coefficient α of the volume and the enhancement coefficient γ of the search frequency have a great influence on the performance of the algorithm. How to effectively balance the algorithm's optimization accuracy and convergence speed, the key is to set the values of parameters α and γ reasonably. In the simulation experiments based on

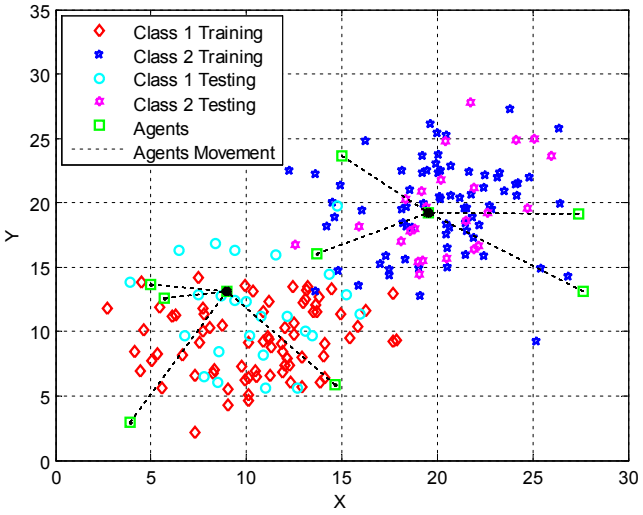
the basic code of BA applied in the data clustering [17], the parameters α and γ can be adjusted in the regular parameter scope. Therefore, this paper mainly adjusts α and γ in the bat algorithm within the allowable modification range so as to achieve the purpose of changing the loudness and velocity of the bat's transmitted pulse, and then the optimal setting of the parameters can be achieved.

A. Change the Single Variable A_i

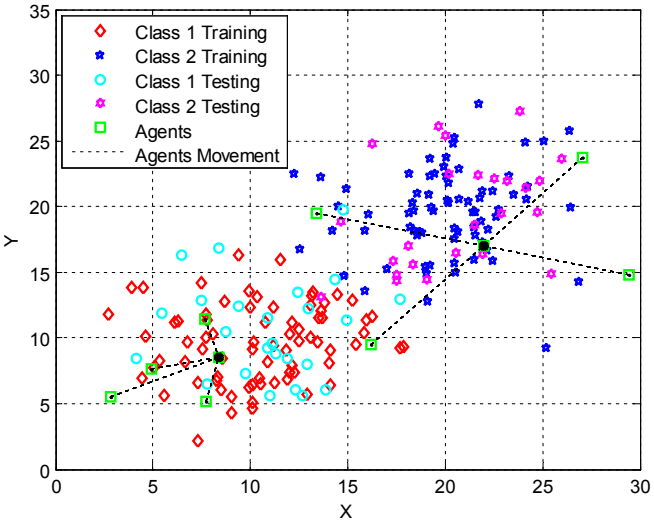
The variable A_i is to control the pulse loudness of the bat. Therefore, the parameter α is adjusted in the adjustable range to carry out simulation experiments, and the bat algorithm is applied to the clustering problem. The simulation result of changing a single variable A_i is shown in Fig. 1. The performance statistic results are listed in Table 1. According to the above simulation figures and table, for the variable A_i of the loudness of the bat's transmitted pulse, the accuracy of the simulation results approaches 100% when the value of parameter α is in the range of [0.3, 0.5]. When α =0.4, the optimization effect of clustering is better and the clustering accuracy stays at 100%. But when α =0.6, the clustering optimization effect is not very good. Therefore, when the value of α is within the range of [0.3, 0.5], the optimization effect of clustering is better.



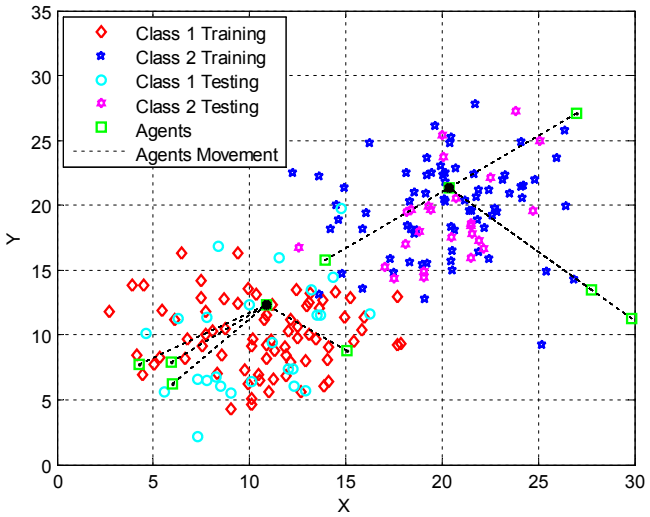
(a) α =0.1, γ =0.9



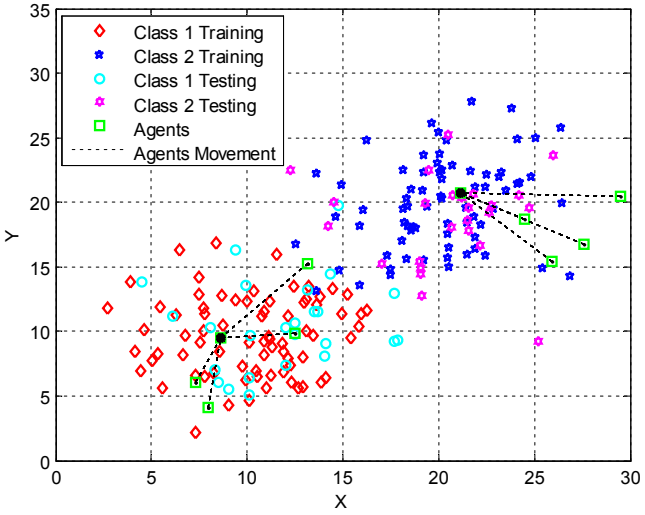
(b) α =0.3, γ =0.9



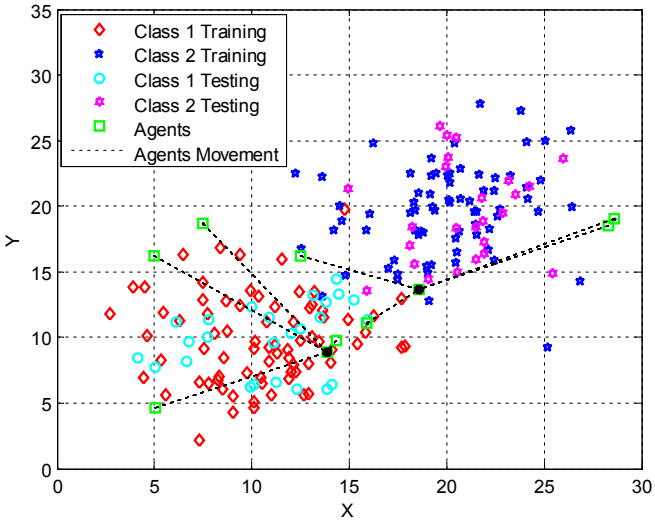
(c) α =0.4, γ =0.9



(d) α =0.5, γ =0.9



(e) α =0.6, γ =0.9



(f) α =0.7, γ =0.9

(g) α =0.9, γ =0.9

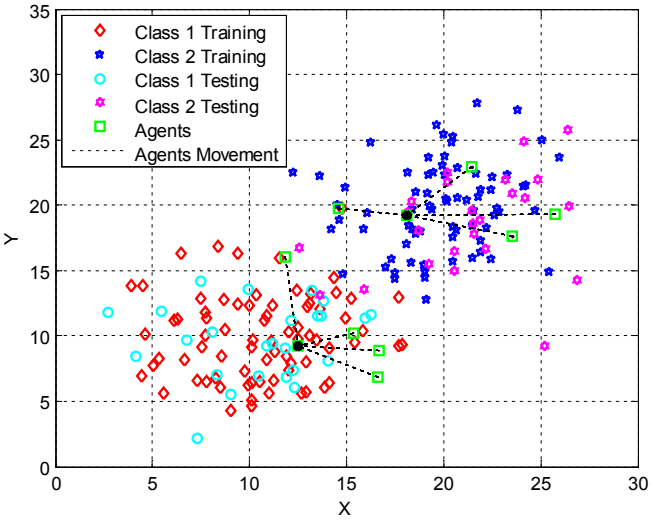
Fig. 1 Simulation results of changing the loudness of bat emission pulse (γ =0.9).

TABLE 1 STATISTICS RESULTS OF SIMULATION PERFORMANCE (γ =0.9)

α	Confusion matrix	Best centers	$f_{\min}$	Time (s)	Overall accuracy (%)
$\alpha=0.1$	24 4 1 21	11.558619 22.636879 9.589000 16.835957	522.2000 380.5079	20.9548	98.00
$\alpha=0.3$	24 1 1 24	8.982723 19.506235 13.161275 19.206444	425.1672 323.8929	24.8250	96.00
$\alpha=0.4$	22 1 3 24	8.387032 21.992719 8.562369 17.035626	342.9907 372.2634	24.6119	92.00
$\alpha=0.5$	24 3 1 22	10.875630 20.346543 12.291792 21.343692	338.1764 336.1867	25.0469	92.00
$\alpha=0.6$	23 0 2 25	8.625638 21.115928 9.457350 20.751380	323.8824 330.4970	24.5106	96.00
$\alpha=0.7$	22 0 3 25	13.854748 18.563990 8.853025 13.693512	387.1243 534.0238	24.6953	94.00
$\alpha=0.9$	23 0 2 25	8.482624 20.866656 12.927213 20.022257	402.1948 303.6602	25.5567	96.00

B. Change the Single Variable r_i

The variable γ is adopted to controls the rate r_i of the emitted pulse. Therefore, the parameter γ can be adjusted within the adjustable range for simulation experiments, and the bat algorithm is applied to the clustering problem. The simulation results of changing a single variable r_i are shown in Fig. 2, and the performance statistic results are listed in Table 2. According to the above simulation results, in the case of α =0.9, the clustering effect of parameter γ in the interval [0.5,10] is good. Obviously, there is a certain difference between the optimization effect obtained by a single change of γ and the optimization effect of a single change of α . Therefore, we will discuss the optimization results in the case when both α and γ are changed to determine the optimal value of the bat algorithm applied in the cluster problem.

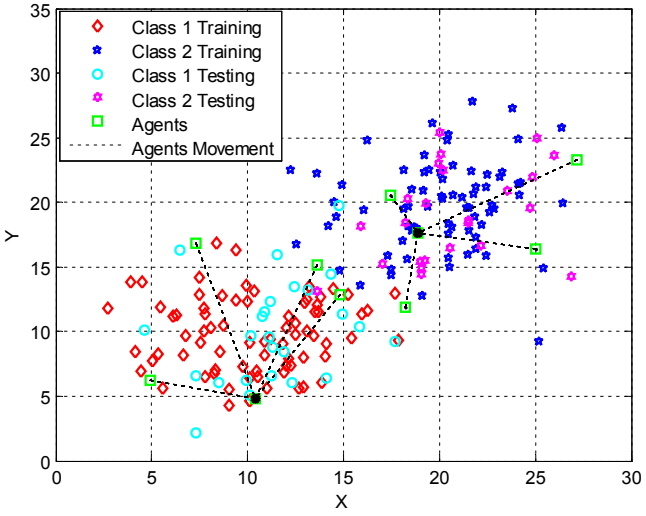


(a) α =0.9, γ =0.1

(b) α =0.9, γ =0.2

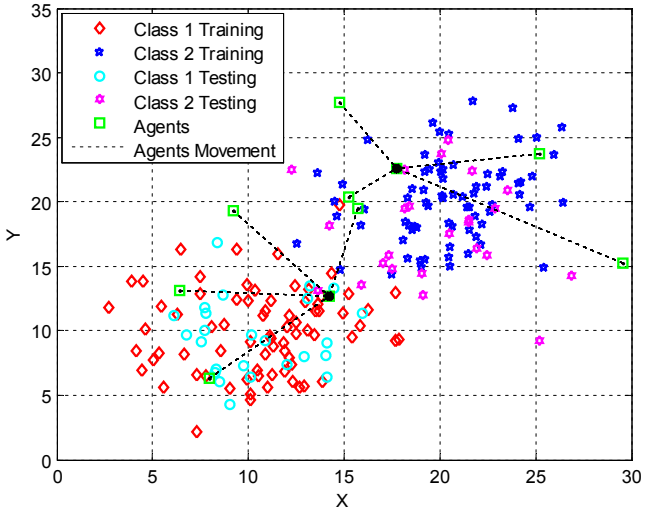
(e) α =0.9, γ =2

(c) α =0.9, γ =0.5



(f) α =0.9, γ =5

(d) α =0.9, γ =1



(g) α =0.9, γ =10

(h) α =0.9, γ =50

(i) α =0.9, γ =100

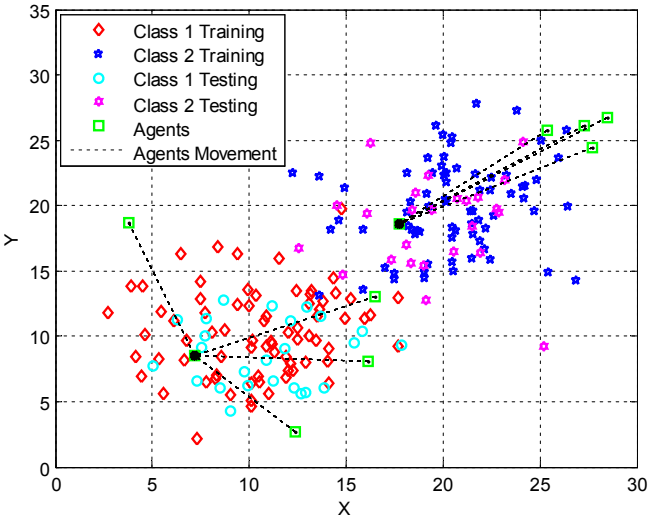
Fig. 2 Simulation results of changing bat pulse rate (α =0.9).

TABLE 2 STATISTICS RESULTS OF SIMULATION PERFORMANCE (α =0.9)

γ	Confusion matrix	Best centers	$f_{\min}$	Time (s)	Overall accuracy (%)
$\gamma = 0.1$	25 2 0 23	12.514379 18.101690 9.241482 19.182926	335.0693 336.4644	28.4174	96
$\gamma = 0.2$	24 2 1 23	11.503299 16.062152 8.995197 21.854291	300.7395 470.0949	24.9736	94
$\gamma = 0.5$	24 2 1 23	13.455828 21.978825 10.431051 22.793103	382.5098 402.0776	17.4244	94
$\gamma = 1$	25 5 0 20	13.180784 23.573337 11.317372 15.700187	361.0874 468.8239	25.3295	90
$\gamma = 2$	25 6 0 19	13.235268 26.088648 8.547952 21.967195	369.3435 551.0793	25.0707	88
$\gamma = 5$	18 0 7 25	10.411758 18.878185 4.808546 17.672576	455.5814 351.4590	24.1408	86
$\gamma = 10$	25 9 0 16	14.223142 17.732594 12.696074 22.579600	433.3649 394.7681	24.5958	88
$\gamma = 50$	21 0 4 25	15.290183 18.356040 5.997596 18.504760	503.9589 355.2370	24.7957	92
$\gamma = 100$	25 5 0 20	15.549496 22.707707 8.629698 21.209652	463.5774 351.3525	24.7250	90

C. *Change the Variables A_i and r_i*

The influence of the single algorithm parameter has small meaning, so it is necessary to carry out the research on the effect of the two parameters of the bat algorithm when solving the clustering problem. In the following simulation experiments, the values of α and γ in the bat algorithm will be combined within the allowable range of modification, so as to achieve the purpose of changing the loudness A_i and rate r_i of the bat's transmitted pulse, and then the parameters can be optimally initialized. The simulation results of changing algorithm parameters are shown in Fig.3 and the performance statistics results are shown in Table 3.



(a) α =0.3, γ =0.5

(b) α =0.3, γ =1

(c) α =0.3, γ =2

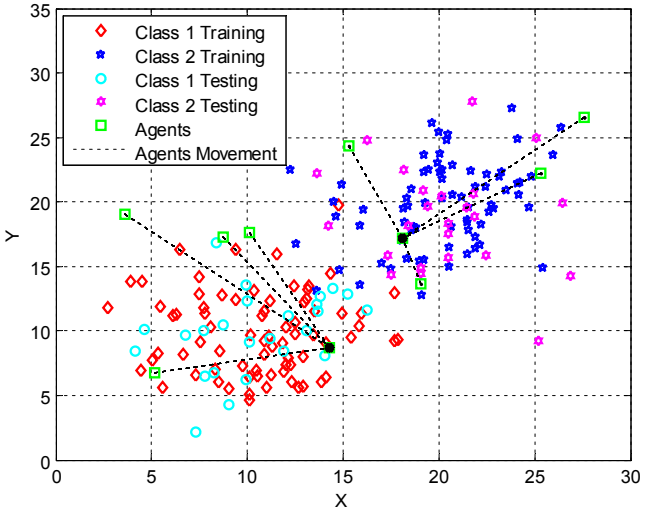
(d) α =0.3, γ =5

(e) α =0.3, γ =7

(f) α =0.3, γ =10

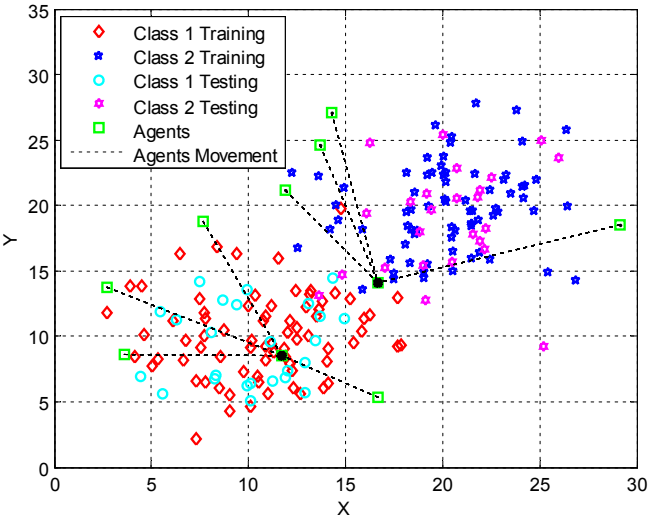
(g) α =0.4, γ =0.5

(h) α =0.4, γ =1



(k) α =0.4, γ =7

(i) α =0.4, γ =2

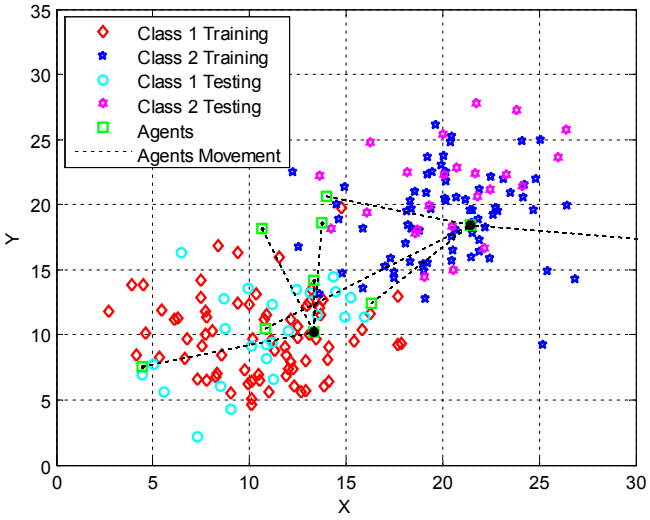


(l) α =0.4, γ =10

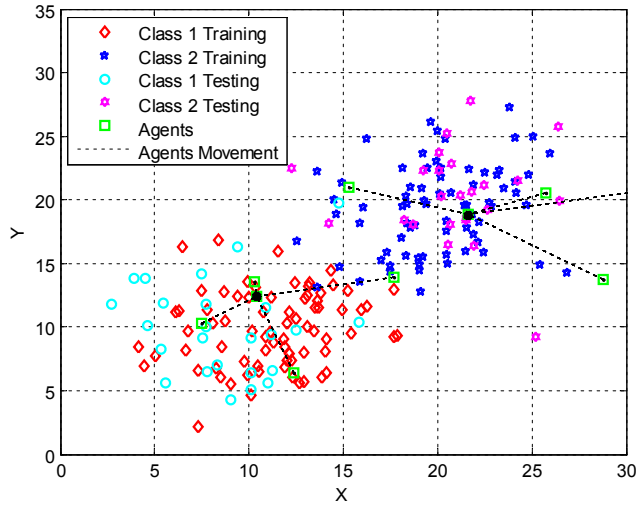
(j) α =0.4, γ =5

(m) α =0.5, γ =0.5

(n) α =0.5, γ =1

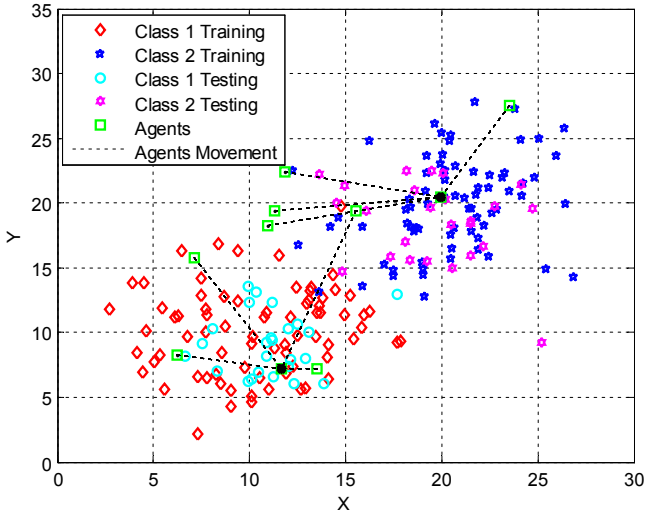


(o) α =0.5, γ =2



(p) α =0.5, γ =5

(q) α =0.5, γ =7



(r) α =0.5, γ =10

Fig. 3 Simulation results of changing the intensity and rate of bat emission pulse.

TABLE 3 COMPARISON OF CLUSTERING ACCURACY AND PERFORMANCE

	α =0.3	α =0.4	α =0.5
$\gamma = 0.5$	90.	94	96
$\gamma = 1$	96	80	82
$\gamma = 2$	88	84	86
$\gamma = 5$	90	90	72
$\gamma = 7$	88	82	80
$\gamma = 10$	64	80	88

According to the obtained simulation results shown in Fig. 3 and Tab. 3, the optimization effect about parameter α is obvious between [0.3, 0.5] alone and the optimization effect about γ is obvious between [0.5, 10]. But a combination of

these two parameters α and γ are not within the optimal range of all optimization effects. It is found that when α is [0.3, 0.5] and γ is [0.5, 5], the optimization effect is as high as 80% to 90%. On the other hand, the optimization effect is higher when γ is [0.5,10] and α is in [0.3, 0.4] and the accuracy of other intervals is obviously decreased. So when we use the bat algorithm in solving the data clustering problem, the parameter γ is initialized in [0.5, 10] and α is initialized in [0.3, 0.4] so as to achieve the best clustering optimization effect.

IV. Conclusion

A new algorithm for solving the clustering problem by combining the foraging behavior of natural bats and the unique flight mode of bats is introduced. Although the bat algorithm can effectively improve the performance of cluster analysis, the reasonable setting of parameters α and γ will effectively accelerate the convergence rate of the algorithm and improve the effective performance of the bat algorithm. In this paper, a large number of simulation experiments are carried out to study the parameter setting of bat algorithm in solving the data clustering problem so as to achieve the better optimization results. In the future, other swarm intelligent optimization algorithm will adopted to solve the clustering problems.

References

[1] Z. C. Zhang, "Determining of soil water stress threshold value with optimum partitioning clustering method," *Journal of Irrigation & Drainage*, vol. 23, no.5, pp. 29-31, 2004.

[2] B. Oviedo, S. Moral, A. Puris, "A hierarchical clustering method: Applications to educational data," *Intelligent Data Analysis*, vol. 20, no.4, pp. 933-951, 2016.

[3] J. Jang, Y. Lee, S. Lee, D. Shin, D. Kim, and H. Rim, "A novel density-based clustering method using word embedding features for dialogue intention recognition," *Cluster Computing*, vol. 19, no.4, pp. 1-12, 2016.

[4] S. Sarmah, D. K. Bhattacharyya, "A grid-density based technique for finding clusters in satellite image," *Pattern Recognition Letters*, vol. 33, no.5, pp. 589-604, 2012.

[5] M. Hayes, Y. S. Pyon, and J. Li, "A model-based clustering method for genomic structural variant prediction and genotyping using paired-end sequencing data," *Plos One*, vol. 7, no.12, pp. 5806-5819, 2012.

[6] C. H. Yang, C. J. Hsiao, and L. Y. Chuang, "Linearly decreasing weight particle swarm optimization with accelerated strategy for data clustering," *Iaeng International Journal of Computer Science*, vol. 37, no.3, pp. 234-241, 2010.

[7] A. Elkamel, M. Gzara, H. Ben-Abdallah, "A bio-inspired hierarchical clustering algorithm with backtracking strategy," *Applied Intelligence*, vol. 42, no.2, pp. 174-194, 2015.

[8] Y. Liu, K. Chen, X. Liao, and W. Zhang, "A genetic clustering method for intrusion detection," *Pattern Recognition*, vol. 37, no.5, pp. 927-942, 2004.

[9] F. Yang, T. Sun, C. Zhang, "An efficient hybrid data clustering method based on K-harmonic means and particle swarm optimization," *Expert Systems with Applications*, vol. 36, no.6, pp. 9847-9852, 2009.

[10] X. S. Yang, S. Deb, S. Fong, X. He, and Y. X. Zhao, "From swarm intelligence to metaheuristics: nature-inspired optimization algorithms," *Computer*, vol. 49, no.9, pp. 52-59, 2016.

[11] X. S. Yang, "A new metaheuristic bat-inspired algorithm," *Computer Knowledge & Technology*, no.284, pp. 65-74, 2010.

[12] A. H. Gandomi, X. S. Yang, A. H. Alavi, and S. Talatahari, "Bat algorithm for constrained optimization tasks," *Neural Computing & Applications*, vol. 22, no.6, pp. 1239-1255, 2013.

[13] S. Kashi, A. Minuchehr, N. Poursalehi, and A. Zolfaghari, "Bat algorithm for the fuel arrangement optimization of reactor core," *Annals of Nuclear Energy*, no.64, pp. 144-151, 2014.

[14] B. Bahmani-Firouzi, R. Azizipanah-Abarghooee, "Optimal sizing of battery energy storage for micro-grid operation management using a new improved bat algorithm," *International Journal of Electrical Power & Energy Systems*, vol. 56, no.3, pp. 42-54, 2014.

[15] T. Niknam, R. Azizipanah-Abarghooee, M. Zare, and B. Bahmani-Firouzi, "Reserve constrained dynamic environmental/ economic dispatch: a new multiobjective self-adaptive learning bat algorithm," *IEEE Systems Journal*, vol. 7, no.4, pp. 763-776, 2013.

[16] M. H. Khooban, and T. Niknam, "A new intelligent online fuzzy tuning approach for multi-area load frequency control: self adaptive modified bat algorithm," *International Journal of Electrical Power & Energy Systems*, no.71, pp. 254-261, 2015.

[17] J. Senthilnath , S. Kulkarni, J. Benediktsson, and X. Yang, "A novel approach for multispectral satellite image classification based on the bat algorithm," *IEEE Geoscience & Remote Sensing Letters*, vol. 13, no.4, pp. 599-603, 2016.

Ling-Feng Zhu is received her B. Sc. degree from Liren College of Yanshan University in 2018. She is currently a master student in School of Electronic and Information Engineering, University of Science and Technology Liaoning, China. Her main research interest is modeling methods of complex process and intelligent optimization algorithms.

Jie-sheng Wang received his B. Sc. And M. Sc. degrees in control science from University of Science and Technology Liaoning, China in 1999 and 2002, respectively, and his Ph. D. degree in control science from Dalian University of Technology, China in 2006. He is currently a professor and Master's Supervisor in School of Electronic and Information Engineering, University of Science and Technology Liaoning.