Data Classification Model of Improved Swarm Intelligence Algorithm Based on Big Data

Dingsheng Deng

Corresponding author's e-mail: dds0904@scun.edu.cn

School of Science and Technology, Sichuan Minzu College, Kangding Sichuan, 626001, China

Abstract: In the era of massive data, people have an urgent need for technology that can automatically and intelligently transform data into useful knowledge. This demand promotes the rapid development of data mining technology. Data classification has been extensively studied in the fields of artificial intelligence, network finance, pattern recognition, and machine learning, and numerous classification modeling algorithms have been produced. Although data classification has made certain breakthroughs in theory and technology, it still has some problems, including: the accuracy and effectiveness of classification modeling algorithms, and the intelligibility of classification rules. This paper introduces the representative ant colony algorithm (ACA) and particle swarm optimization algorithm (PSOA) in the swarm intelligence algorithm into data classification mining, and proposes a data classification model based on the improved swarm intelligence algorithm of big data.

Keywords: Big Data, Intelligent Algorithm, Data Classification, Ant Colony Algorithm

1. Introduction

The rapid development of data collection technology has made it easy for people to obtain a large amount of data, but it has been difficult to obtain the hidden knowledge and internal relationships behind the large amount of data by using traditional data processing and statistical analysis tools [1-2]. People's urgent need for technology that can automatically and intelligently transform data into useful knowledge has promoted the rapid development of data mining [3-4]. How to extract accurate, understandable, and usable knowledge and information from massive data has become one of the hot issues in the field of data mining today [5-6].

At present, the research on data classification problems has achieved some results, such as decision tree classification algorithm, neural network classification algorithm, Bayesian classification algorithm, genetic classification algorithm, rough set classification algorithm and support vector machine and other classification algorithms [7]. However, each algorithm has its own shortcomings.

This paper mainly uses the improved ACA and PSOA based on big data to study the classification problem of data mining, and completes the design of a feasible learning model construction algorithm that can better solve the classification problem. In-depth research provides scientific and feasible modeling and data analysis methods.

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334269

2. Data Classification Model of Improved Swarm Intelligence Algorithm Based on Big Data

2.1 Improved Particle Swarm Optimization Data Classification Modeling Algorithm Based on Big Data

(1) PSOA

In PSO, each particle represents a possible solution, and the entire population realizes the search for the optimal solution in a multi-dimensional space through mutual competition and cooperation. In the D-dimensional space, each particle is regarded as a solution, consisting of the current position $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ and the current velocity $\mathbf{V}_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$. For the global PSO, the particle $\boldsymbol{X}_i$ is updated iteratively according to its historical optimal value $\mathbf{P}_i = (p_{i1}, p_{i2}, \ldots, p_{iD})$ and the global optimal value $\mathbf{P}_g = (p_{g1}, p_{g2}, \ldots, p_{gD})$. Update position and speed of each particle:

$$v_{id}^{t+1} = v_{id}^{t} + c_1 \cdot r_{1id} \cdot (p_{id}^{t} - x_{id}^{t}) + c_2 \cdot r_{2id} \cdot (p_{pd}^{t} - x_{id}^{t}) \tag{1}$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{2}$$

Where t is the current evolutionary algebra; c_1 and c_2 are two positive acceleration constants; r_1 and r_2 are two random numbers between [0, 1]. In each iteration, the fitness value $f(\mathbf{X}_i)$ of the particle determines the performance of the particle. Shi and Eberhart proposed the inertial weighting factor ω,

$$v_{id}^{t+1} = w \cdot v_{id}^{t} + c_1 \cdot r_{1id} \cdot (p_{id}^{t} - x_{id}^{t}) + c_2 \cdot r_{2id} \cdot (p_{gd}^{t} - x_{id}^{t}) \tag{3}$$

The introduction of the constraint factor χ is to control the velocity of the particles:

$$v_{id}^{t+1} = c \left(v_{id}^{t} + c_1 \cdot r_{1id} \cdot (p_{id}^{t} - x_{id}^{t}) + c_2 \cdot r_{2id} \cdot (p_{gd}^{t} - x_{id}^{t})\right) \tag{4}$$

$$c = \frac{2}{\left|2 - f - \sqrt{f^2 - 4f}\right|}, \quad f = c_1 + c_2 \tag{5}$$

Where $\phi>4$, when the inertia weight factor $\omega=0.7298$, $c_1 = c_2 = 1.49445$, it is equivalent to taking $\phi=4.1$ in the particle swarm algorithm using the shrinkage factor, so the shrinkage factor $\chi=0.729$, and the particle swarm optimization at this time the algorithm is convergent.

According to the object information referenced by the particles in the update process, the classic model of the PSOA can be divided into: a global model and a local model.

In the global model, each particle shares information with other particles in the population and moves to the best historical position in the population. Therefore, the algorithm converges faster, and the problem is that the algorithm is more likely to fall into the local optimum.

In order to avoid the local optimal problem of the global model, the particles in the local model only move to the optimal position in their neighborhood, so the optimal value of different particles is different, thereby increasing the diversity of the population. But because different particles move to different optimal values, the convergence speed of the algorithm decreases. The neighborhood structure includes pyramid, star, small neighborhood, and von Neumann structure. The study found that for complex problems, the advantage of the small neighborhood structure is more obvious, while for simple problems, the large neighborhood structure is relatively better.

The algorithms proposed in this paper are based on global models.

(2) Orthogonal PSOA based on quadratic interpolation

First, we give the flow of the entire algorithm, and then discuss the three operators introduced separately.

In order to prevent particles from exceeding the search space, the following constraints are set: only when the current position and velocity of the particles are within the constraints, the historical and global optimal values of the particles are updated, and Gap_{max} is set to 5. After repeated experiments, we set the maximum threshold of velocity in each dimension as $V_{max}^{d} = 0.2 \times X_{max}^{d}$, and the minimum threshold value is $V_{min}^{d} = 0.2 \times X_{min}^{d}$.

1) Quadratic interpolation operator

As a mathematical optimization method, quadratic interpolation has been applied to evolutionary optimization algorithms. For example, it is used to initialize the population in the differential evolution algorithm, and it is also used as a non-linear crossover operator to generate a new population. In the algorithm, quadratic interpolation (QI) is used to make full use of the entire population information.

After constructing $\mathbf{P_O} = (P_{o1}, P_{o2}, \ldots, P_{oD})$, the new orthogonal learning mechanism can be applied, and the corresponding speed update formula is adjusted as follows:

$$v_{id}^{t+1} = w' v_{id}^{t} + c' r_{id}' (p_{od}^{t} - x_{id}^{t}) \tag{6}$$

$$\mathbf{P_O} = \mathbf{P_i} \text{ Å } \mathbf{P_n} \tag{7}$$

In the optimization process, each particle must obtain corresponding empirical information from its historical optimal value and global optimal value for iterative update. In this way, the entire population will contain a wealth of information about the given problem. Therefore, $\mathbf{P}_i$ interacts with the entire population to generate a better vector $\mathbf{P}_i' = (P_{i1}', P_{i2}', \ldots, P_{iD}')$. For each particle, take its historical optimal value $\mathbf{P}_i$ and two arbitrary particles $\mathbf{X}_b$, $\mathbf{X}_c$ to construct a quadratic curve, and take the lowest point of the curve as the historical optimal value $\mathbf{P}_i'$ of the next generation of particles.

$$p_{id}^{'} = \frac{1}{2} \frac{(\mathbf{x}_{bd}^{2} - \mathbf{x}_{cd}^{2}) f(\mathrm{P_i}) + (\mathbf{x}_{cd}^{2} - p_{cd}^{2}) f(\mathrm{X_b}) + (p_{id}^{2} - \mathbf{x}_{bd}^{2}) f(\mathrm{X_c})}{(\mathbf{x}_{bd} - \mathbf{x}_{cd}) f(\mathrm{P_i}) + (\mathbf{x}_{cd} - p_{id}) f(\mathrm{X_b}) + (p_{id} - \mathbf{x}_{bd}) f(\mathrm{X_c})} \tag{8}$$

In order to avoid that the constructed quadratic curve is not concave downwards, after obtaining the extreme value $f(\mathbf{P}_i')$, compare it with the historical optimal value $f(\mathbf{P}_i)$ of the particle, if $f(\mathbf{P}_i')$ is less than $f(\mathbf{P}_i)$, replace $\mathbf{P}_i$ with $\mathbf{P}_i'$, and update the fitness value at the same time.

2) Reverse learning operator

The Operator Backward Learning (OBL) was proposed by Tizhoosh and used as an initialization population to improve the convergence speed of the algorithm and the accuracy of the solution.

In the PSO optimization process, the empirical information of the particles can be memorized and used to update the next generation of particles. The worst particle in the population has the largest difference between its fitness value and the optimal value, and its impact on the population will be more obvious. First give some definitions related to OBL: Let $\mathbf{X}(x_1, x_2, \ldots, x_D)$ be a point in a D-dimensional space, where $x_1, x_2, \ldots, x_D \in R(x_i \in [a_i, b_i] \forall i \in \{1, 2, \ldots, D\})$. The reverse point can be defined as follows:

$$\breve{x}_i = a_i + b_i - x_i \tag{9}$$

The worst particle in the population is defined as $\mathbf{X}_{\text{worst}}(x_{\text{wost1}}, x_{\text{worst2}}, \ldots, x_{\text{worstD}})$, and the element on the D dimension of the reverse particle $\breve{\mathbf{X}}_{\text{worst}}$ is $\breve{x}_{\text{worst}} = X_{max} + X_{min} - x_{\text{wost}}$, X_{max} and X_{min} are respectively the search space Upper and lower limits.

In some cases, when the population falls into the local optimum, the fitness value of $\breve{\mathbf{X}}_{\text{worst}}$ will be worse than that of $\mathbf{X}_{\text{worst}}$ according to the principle of reverse learning. However, through this reverse processing, particles in a new direction will be obtained, which is opposite to the original direction, and finding such a new direction is exactly the purpose of citing the reverse learning operator.

3) Elite selection operator

In the PSOA based on orthogonal learning, although $\mathbf{P}_g$ is no longer an independent factor in the particle update, it is still a key element of $\mathbf{P_0}$ in the update formula, and it plays an important role in guiding the algorithm to converge. However, it is the only solution without a guiding vector. Especially when dealing with complex and multimodal functions, it is difficult to get rid of the local optimum in the convergence stage. Therefore, we introduce an elite selection strategy to make the entire population converge to a better position.

Although the population falls into the local optimum, most dimensions of the global optimum solution still have a lot of useful information, so we only apply the elite selection mechanism to a certain dimension of $\mathbf{P}_g$, and keep the parameters in other dimensions. One dimension is selected arbitrarily for perturbation operation, and the probability of each dimension being selected is the same. The formula is as follows:

$$P^d = P^d + (X_{\max} - X_{\min}) \times N(\mu, \sigma^2) \tag{10}$$

P^d is the value corresponding to the dth dimension of $\mathbf{P}_g$; $[X_{max}, X_{min}]$ is the search interval of the problem to be solved. In order to make the newly generated point lie in the neighborhood of the original P^d, the mean value μ of $N(\mu, \sigma^2)$ is taken as 0, and the standard deviation

decreases linearly with the algebra.

$$s = s[\max] - (s[\max] - s[\min]) \times \frac{gen}{Gap[\max]} \tag{11}$$

3. Experimental Setup

3.1 Based on the Improved Ant Colony Data Classification Modeling Algorithm Experiment

(1) Setting and preprocessing of data set

Twelve standard data sets were selected from the UCI machine learning database as experimental data sets to evaluate the performance of the proposed colony data classification modeling algorithm based on big data improvements.

The data sets selected for the experiment were Anneal, Cancer-Breast, Credit-a, Dermatology, Glass, Heart-c, Heart-h, Ionosphere, Liver-disorder, and Pima, testing algorithm classification problems. A total of 10 data sets to do. Attributes that include binary and multivariate classification problems include discrete and continuous attribute values, and these data sets are used to test related classification algorithms. Table 1 shows the main characteristics of the data set used in the experiment. Take the breast-cancer data set as an example: it has a total of ten attributes, namely age (age), menopause (menopause), tumor size (tumorsize), number of invaded lymph nodes (inv-nodes), and presence or absence of nodules (Node-caps), degree of malignancy (deg-malig), location of the tumor (breast), quadrant of the tumor (breast-quad), whether it is irradiat (irradiat), and whether it recurs (class). There are two types of recurrence (class), which are recurrence-events and non-recurrence-events.

Table 1. Data set characteristics

	Number	Number of attributes	Number of categories
Anneal	890	39	5
Breast-cancer	278	9	2
Credit-a	684	15	2
Dermatology	365	38	5
Glass	218	9	8
Heart-c	305	12	5
Heart-h	298	12	5
Ionosphere	356	35	2
Liver-disorders	349	6	2
Pima	798	7	2

Since this algorithm is suitable for the classification of discrete attributes, continuous attributes must be discretized through preprocessing steps. The discretization process in this article uses the C4.5_Disc discretization method, which is implemented using the well-known C4.5 algorithm. The discretized data set contains only two attributes: the discretized attribute and category.

(2) Improved ant colony classification algorithm parameter settings

There are six parameters in the improved ant colony classification algorithm (Ant-MinerPAE) that need to be set: the maximum number of iterations m, the number of ants colony-size, the pheromone volatilization coefficient ρ, the minimum number of instances covered by each rule min_cases_per_rule, the largest in the training set the number of uncovered instances is maximum uncovered, and the number of identical rules no_rules_converge for judging whether the ant has converged.

Ant-MinerPAE's parameter settings are not optimized, but even if the above parameter settings are used, the algorithm can get better results. The parameters in other algorithms use the default values used by the authors in the relevant literature, and these values are better choices used in their research.

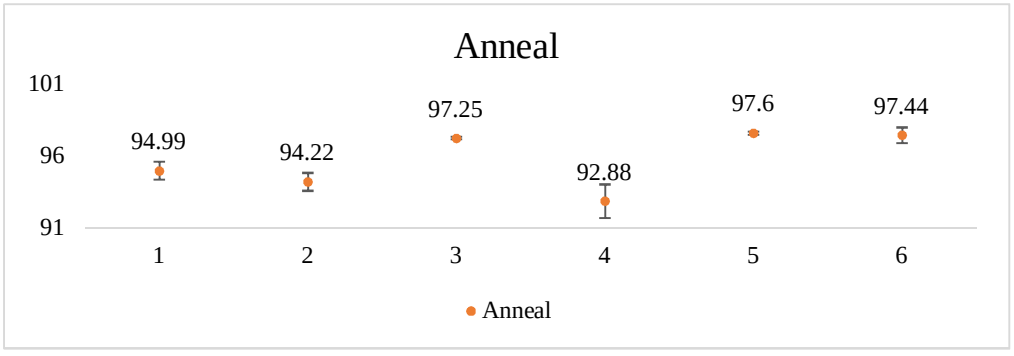
3.2 Research Experimental Settings Based on the Improved Particle Swarm Optimization Data Classification Modeling Algorithm Based on Big Data

We compared it with five classification algorithms: CLPSO, FDRPSO, r3PSO, LIPS and OLPSO-G. In this article, we will select four representative functions from the four test modes F1, F7, f6 and F25. In the experiment, for fair comparison, the termination conditions of all algorithms are: the maximum number of function evaluations (FEs) is 300,000 times, the population size is 40, each particle is a 30-dimensional individual, and all test results are run 25 times average value. Other parameters are set as follows: CLPSO, FDRPSO, LIPS, OLPSO-G and QIOLPSO-G inertia weighting factors are all linearly decreasing with algebra between [0.4, 0.9]. Gap_{max}, acceleration coefficient c, maximum and minimum speed thresholds V_{max}^{d} and V_{min}^{d} are all values obtained through trial and error.

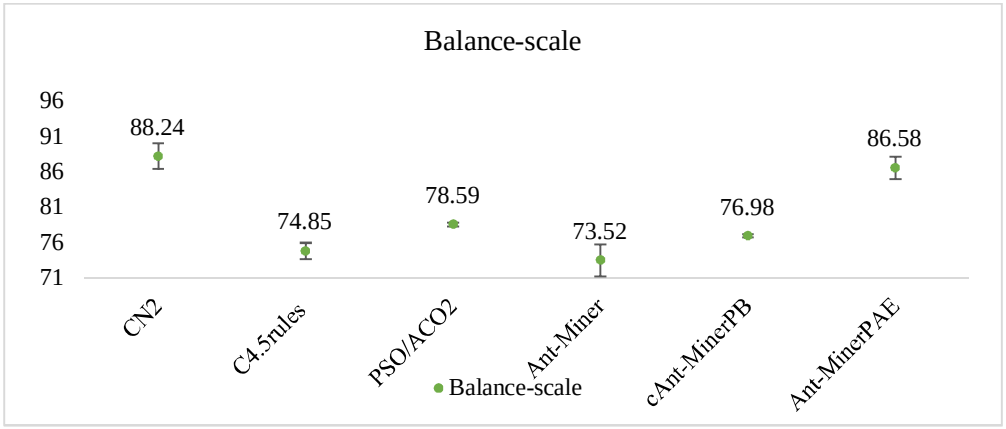
4. Research and Experimental Analysis on Data Classification Model of Improved Swarm Intelligence Algorithm Based on Big Data

4.1 The average prediction accuracy of the improved ant colony classification algorithm

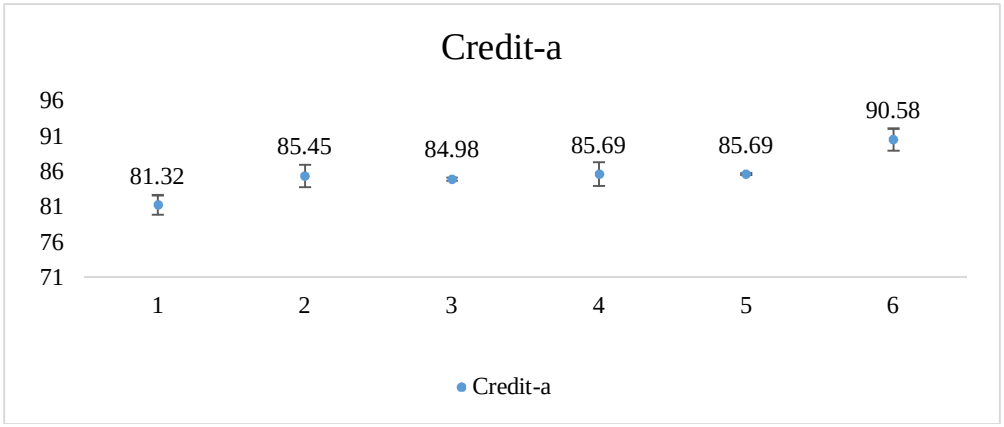
Figure 1 is a comparison chart of the prediction accuracy of 6 classification algorithms on different data sets, including CN2, C4.5rules, PSO/ACO2, Ant-Miner, cAnt-MinerPB, Ant-MinerPAE (Algorithm 1-6).



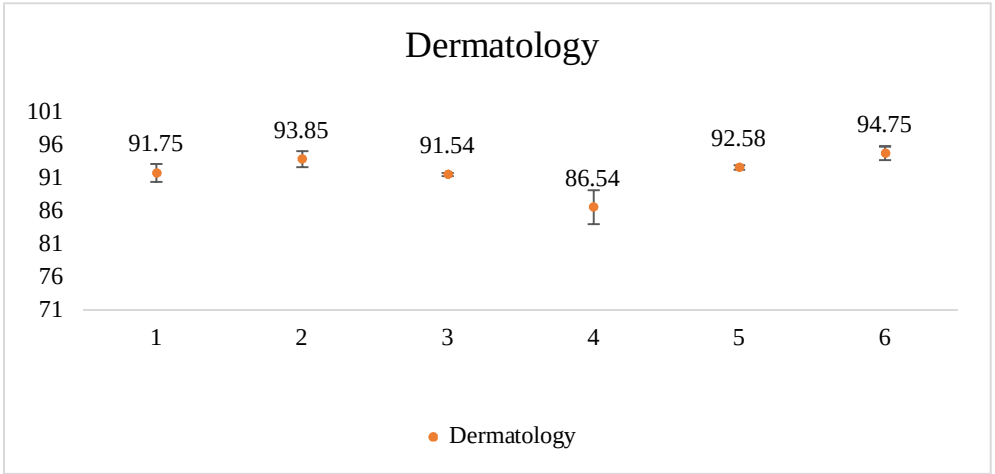
(a) Anneal data set



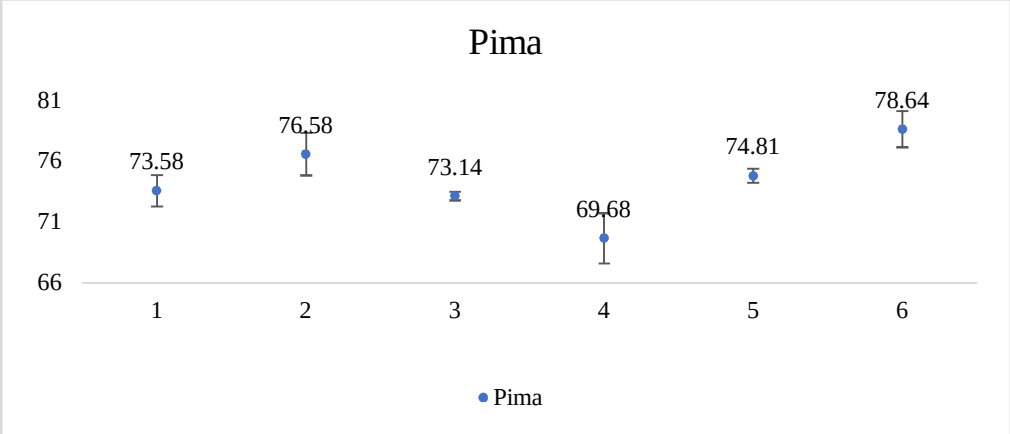
(b) Balance-scale data set



(c)Credit-a data set



(d) Dermatology data set



(e) Pima data set

Figure 1. Comparison of average prediction accuracy of 6 algorithms

In the classification experiment of the Anneal data set, the result of the cAnt-MinerPB algorithm is the best, which is 97.6±0.1. In the classification of the Balance-scale data set, CN2 has the highest prediction accuracy of 88.24±1.81, which is better than the results of several other algorithms. On the Credit-a data set, the prediction accuracy of the Ant-MinerPAE algorithm reached 90.58±1.58, while none of the other algorithms exceeded 90. On the Dermatology data set, the prediction accuracy of the Ant-MinerPAE algorithm reached 94.75±1.06, which was slightly higher than the prediction accuracy obtained by other algorithms. The prediction accuracy of the Ant-MinerPAE algorithm in the Pima data set is 78.64±1.44, which is also higher than other algorithms. Experiments show that among the 5 data sets, the Ant-MinerPAE algorithm has higher prediction accuracy in 3 data sets.

4.2 Average Fitness Function Value of the Improved PSOA

For most functions, OLPSO-G has a certain degree of competition. Therefore, the next experiment is to compare the quadratic interpolation orthogonal PSOA (QIOLPSO-G) and the orthogonal learning-based PSOA (OLPSO-G)). At the same time, four representative functions were selected for testing from four types of functions: F1, F7, f6 and F25. First, we test the diversity of the population. This standard is called the overall standard deviation, which is recorded as PSD. The larger the PSD value, the higher the diversity of the population; and vice versa. This article only selects the 9th generation data to show as follows:

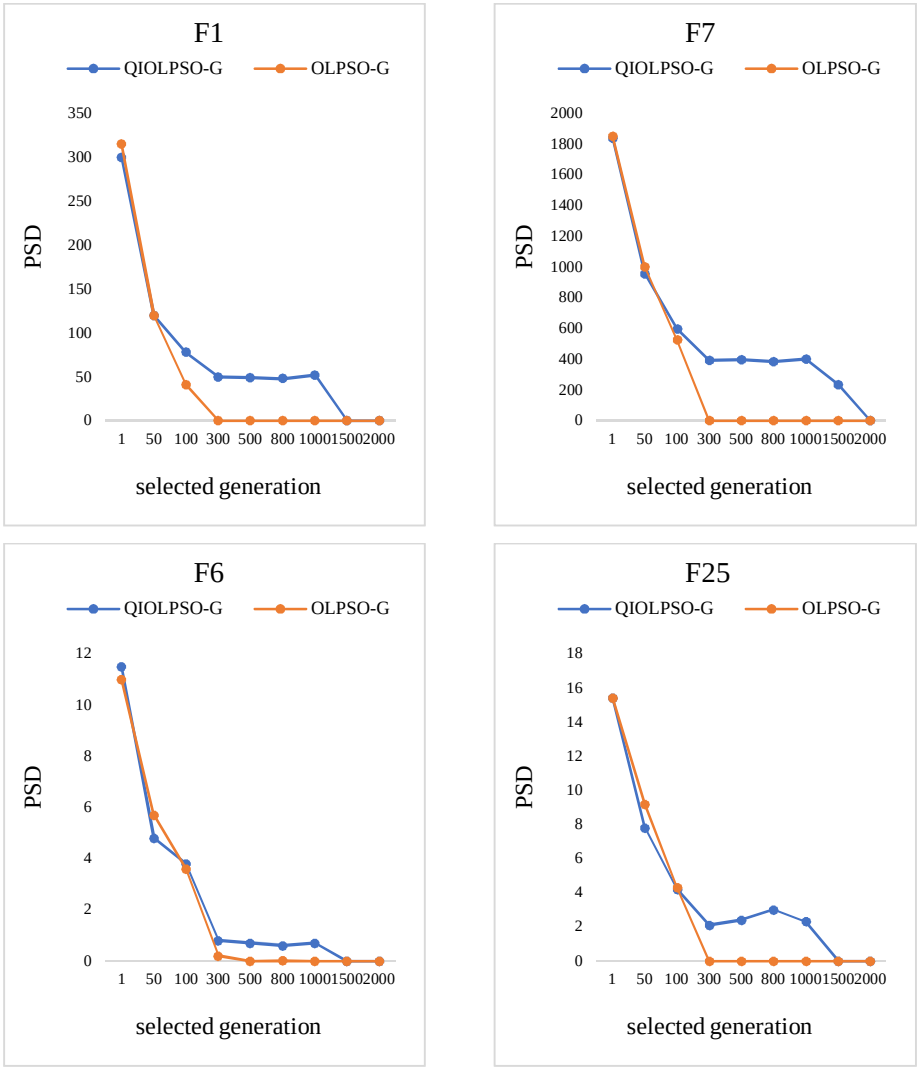


Figure 2. Overall standard deviation

From Figure 2 we can see that the psd value obtained by the QIOLPSO-G algorithm is higher than that of the OLPSO-G, which proves that our improved algorithm is more diverse than the original algorithm. In the optimization process, as the algorithm gradually converges, its diversity will gradually decrease. However, the diversity of QIOLPSO-G has always been higher than that of OLPSO-G. The maintenance of diversity has an important impact on processing optimization problems, especially complex functions. The higher the diversity, the closer the final solution will be to the actual optimal value of the function.

In order to make it more difficult to clearly show the role of the elite selection strategy operator, we list the corresponding results as follows:

Table 2. Average fitness function values corresponding to different improvement mechanisms

	1	2	3	5	5	f(x*)
F1	-450±0	-450±0	-450±0	-450±0	-450±0	-450
F7	-159.99 ±6.8e-03	-159.99 ±8.5e-03	-159.99 ±0.03	-159.99 ±0.03	-159.99 ±6.58e-03	-160
F25	460 ±1.25e-12	460 ±6.58e-11	460 ±2.38e-12	460 ±0	460 ±0	250
F6	4.49e-15 1.52e-15	5.94e-15 ±1.95e-15	6.98e-15 ±1.91e-15	5.64e-15 ±2.58e-15	4.28e-15 ±1.58e-15	0

Only QI: only quadratic interpolation; only OBL: only reverse learning operator; only Gau mu: only Gau mu; OBL+QI: reverse learning operator and quadratic interpolation; Gau mu+OBL+QI: Gaussian mutation, reverse learning operator and quadratic interpolation.

It can be seen from Table 2 that the result of the algorithm has indeed been improved correspondingly after adding the elite selection strategy.

The three operators proposed in this paper are easy to implement, but will require additional function evaluation times. But when the number of evaluations is the same, our improved algorithm has greater advantages than the comparison algorithm, especially in processing complex multi-modal functions. It shows that our algorithm is more diverse, more convergent, and capable of jumping out of the local optimum.

5. Conclusions

The continuous improvement and development of data mining technology can precisely help people find potentially useful information and knowledge from a large amount of noisy and fuzzy data, and classification is one of the more important technical methods. Therefore, the key to the problem is to explore more effective data classification methods. In recent years, improved swarm intelligence algorithms based on big data have become the focus of research by scholars due to their simple concepts, fewer parameters to be adjusted, and easy programming to achieve. Therefore, the classification algorithm optimized by swarm intelligence has also become a hot spot in the classification field.

ACKNOWLEDGMENT ： This paper is supported by the key project of natural science of Sichuan Minzu College (No. XYZB2201JG) and Science and technology plan project of Ganzi Science and Technology Bureau (No. 22kjjh0013).

References

[1] Kim J K, Mi J R, Lee J S, et al. Improved Prediction of the Pathologic Stage of Patient with Prostate Cancer Using the CART–PSO Optimization Analysis in the Korean Population[J]. Technology in Cancer Research & Treatment, 2017, 16(6):740-748.

[2] Ibrahim R A, Ewees A A, Oliva D, et al. Improved salp swarm algorithm based on particle swarm optimization for feature selection[J]. Journal of ambient intelligence and humanized computing, 2019, 10(8):3155-3169.

[3] Panda N, Majhi S K. Improved spotted hyena optimizer with space transformational search for

training pi-sigma higher order neural network[J]. Computational Intelligence, 2020, 36(1):320-350.
[4] Parvathy G, Bindhu J S. A Probabilistic Generative Model for Mining Cybercriminal Network from Online Social Media: A Review[J]. International Journal of Computer Applications, 2016, 134(14):1-4.
[5] Subbulakshmi B, Deisy C. An improved incremental algorithm for mining weighted class-association rules[J]. International Journal of Business Intelligence and Data Mining, 2018, 13(1-3):291-308.
[6] Zhang X, Wang M. Improved SVM classification algorithm based on KFCM and LDA[J]. Journal of Physics: Conference Series, 2020, 1693(1):012107 (6pp).
[7] Dan S, Kui L, Xufan D. Research on mail classification algorithm based on improved convolutional neural network[J]. Journal of Physics: Conference Series, 2021, 1871(1):012097 (8pp).