YUAN ZHANG
FENG ZHANG
YONG-HENG ZHANG
and
YE ZHANG

Water resources big data classification based on multi-objective optimization for mining area

In order to solve the uncertainty of support vector machine kernel function parameters and solve the optimal selection of kernel parameters in the classification algorithm of big data of mining area water resources, a mining area water big data classification algorithm based on PSO-SVM hybrid optimization is proposed. This algorithm solves the existence of inseparable regions and error accumulation in the support vector machine multi-classification method. Based on the analysis of basic particle swarm optimization algorithm and SVM algorithm working principle, the advantages of PSO and SVM algorithm are mixed, and the convergence speed is moderately improved to make it have the ability of self-adaptation, and the fine search is performed in the final stage. Expand the width and depth of parameter search to meet the characteristics of diversification and concentration. The results show that the hybrid soft calculation method proposed in this paper can improve the accuracy of classification and prediction, and classification accuracy and classification time are significantly improved, and it is an effective multi classification algorithm.

Keywords: Support vector machine, big data classification, soft computing, mining area, particle swarm optimization

1. Introduction

As a classic and important topic in the field of mining area data mining, the classification problem has always been concerned by the academic community[1]. However, with the promotion of the Internet of Things and the advent of the "big data" era, traditional mining area data classification methods are facing new challenges, the first and foremost is the change of data form, from traditional static data to dynamic data flow. Compared to static data, dynamic data has three characteristics, namely, massiveness, real-time and dynamic variability, which greatly increases the difficulty of data stream classification[2].

Therefore, how to design a mining area data stream classification model not only satisfies the characteristics of the data stream, but also can effectively classify the data stream, which has become a hot issue in current academic research.

With the in-depth study of the integration model, the research of the integrated classification model is divided into two parts, namely, the construction of weak classifier and the summary of classification results[3]. For the weak classifier construction part, it is mainly to establish different weak classification models by adjusting the internal parameters of the model and the training set, and provide the classifier basis for the construction of the integrated model[4]. It should be noted that for classifiers, the same classifier may be different, but which method is better, the academic community is still not conclusive, and specific analysis needs to be carried out in practical problems.

In this paper, a data flow classification model based on ensemble learning is proposed to detect the abnormal value of mining area water resources data. An integrated learning model is constructed by combining different kernel functions.

2. Mining area data classification method based on PSO optimization learning

In view of the massive nature of big data streams, classification models are required to have strong learning capabilities to adapt to different data rings. Changes in the environment, but the traditional single model structure classification method is difficult to meet this requirement. In response to this problem, this paper uses PSO learning ideas to construct a classification model to classify data streams. The proposed model is constructed using the support vector machine model and the self-organizing map model[5]. The model parameters are initialized and optimized by genetic algorithm and particle swarm algorithm to achieve the best classification effect.

2.1. Particle swarm optimization algorithm

Particle filter is a method of Monte Carlo simulation based on recursive Bayesian filtering[6]. The key idea is to use a set of weighted sum of the weights associated with a random sample to represent posterior probabilities The basic PSO algorithm can be described as follows: Let the search space be N dimension, the number of particles is P, the position and velocity of the i-th particle in the N-dimensional search

Messrs. Yuan Zhang*, Mathematics and Statistics, Yulin University, Yulin 719000, Shaanxi and Feng Zhang, Yong-Heng Zhang and Ye Zhang, School of Information Engineering, Yulin University, 719000, Yulin, China
*Email of the corresponding author: 54833837@qq.com

space are $X_i = (x_{i1}, x_{i2}, ... x_{iN})$ and $V_i = (v_{i1}, v_{i2}, ... v_{iN})$. The individual extremes and group extremes of the particles are $P_i = (p_{i1}, p_{i2}, ... p_{iN})$ and $P_g = (p_{g1}, p_{g2}, ... p_{gN})$. The particle flight is shown in Fig. 1.

PSO algorithm mathematics are as follows:

Its optimization problem model: *min f(x)*

Let *f(x)* search space is *D*-dimensional, the total number of particles is *N*, The position of the $(_{i = 1, 2, ... N})$ particles is $X_i = (X_{i1}, X_{i2}, ... X_{ij}, ... X_{iD})$, the flight speed of the *i* particles is $V_i = (V_{i1}, V_{i2}, ... V_{ij}, ... V_{iD})$, the optimal position of the *i* particle flight history is *pbest*, then $P_i = (P_{i1}, P_{i2}, ... P_{ij}, ... P_{iD})$, in this group, at least one particle is optimal, denoted *gbest*, then $P_{gbesti} = (P_{gbest1}, P_{gbest2}, ... P_{gbestD})$ is the global history optimal position of the current group. $fitness_i = f(x_i)$ represent the position, velocity and fitness value of *i* particles in *i*.

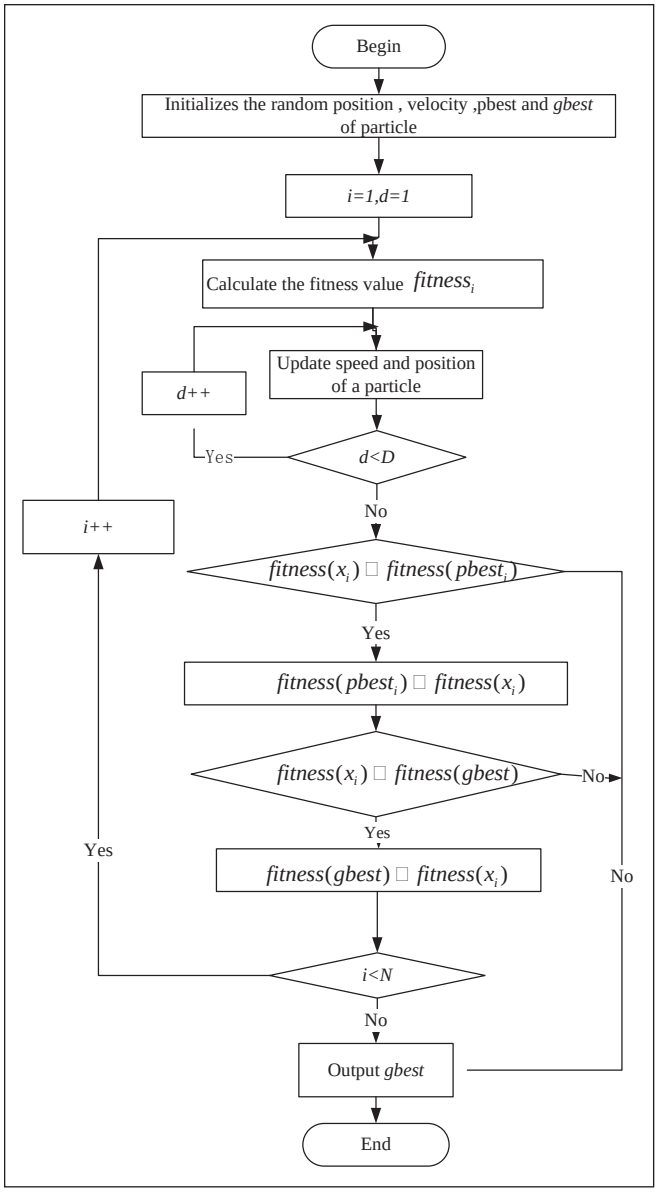

Fig. 1: Particle swarm optimization process

The position update formula of each particle is:

$$v_{ij}(t+1) = \omega.v_{ij}(t) + c_1.r_1.(p_{ij} - x_{ij}(t)) + c_2.r_2.(p_{gbestj} - x_{ij}(t)) \quad (1)$$

$$x_{ij}(t+1) + x_{ij}(t) + v_{ij}(t+1)$$

WWhere, *t* represents the number of iterations, $i = 1, 2, ..., N$; $j = 1, 2, ..., D$; $c_1, c_2 > 0$ factor represents individual learning and social learning factor, r_1 and r_2 are both in the range between [0,1] independent random factor[7]; ω represents the inertia weight used to weigh the ability of local optimum and global optimum capacity. In order to balance global and local search capabilities, and its value should decrease linearly with the evolutionary algorithm, ω is defined as:

$$\omega = \omega_{min} + (iter_{max} - iter) \times (\omega_{max}\text{-}\omega_{min})/iter_{max} \quad (2)$$

Where, ω_{min}, ω_{max} respectively maximum and minimum weight factor, *iter* is the current iteration number, $iter_{max}$ is the total number of iterations. Particle swarm optimization process shown in Fig. 1, the process of the algorithm is as follows:

(1) Random initialization position and velocity of the particle swarm.

(2) Calculate the fitness value of each particle
$fitness_i = f(x_i)$, corresponding initialization
$pbest_i = fitness_i$,
$gbest = \min(fitness_1, fitness_2, ..., fitness_N)$, i = 1, 2, ..., N

(3) For each particle, its fitness compared with *pbest*, if it is the best, it is the best as the current position and update the *gbest* and *pbest*.

(4) The adaptation values of each particle are compared with the adaptation values of *pbest*. If better, then as *gbest*.

(5) Iterative update speed and position of a particle.

(6) If the number of iterations unfinished or find a satisfactory adaptation value, will continue to calculate the fitness value of each particle.

(7) Output *gbest*.

2.2. Support vector machine algorithm

If a linear function can completely separate the samples correctly, they are said to be linearly separable, otherwise called non-linear separable[8]. In the sample space, the hyperplane can be described by the following linear equations.

$$g(x) = w'x + b = 0$$

Assuming that it has completed the separation of the samples and the labels for the two samples are {+1, -1}, then for a classifier, $g(x)>0$ and $g(x)<0$ can be separated. Represents two different categories, which are +1 and -1 respectively[9]. But it is not enough to separate light. The core idea of SVM is to make the maximum effort to make the separate two categories have the maximum interval, which makes the separation more credible. Moreover, there is good classification prediction ability for unknown new samples[10]. In order to describe the data points closest to the separating

hyperplane, we need to find two hyperplanes that are parallel to this hyperplane and are equal in distance:

$$H_1 : y = w^t x + b = +1,\ H_2 : y = w^t x + b = -1 \tag{3}$$

The schematic diagram of hyperplane in SVM is as shown in Fig. 2.

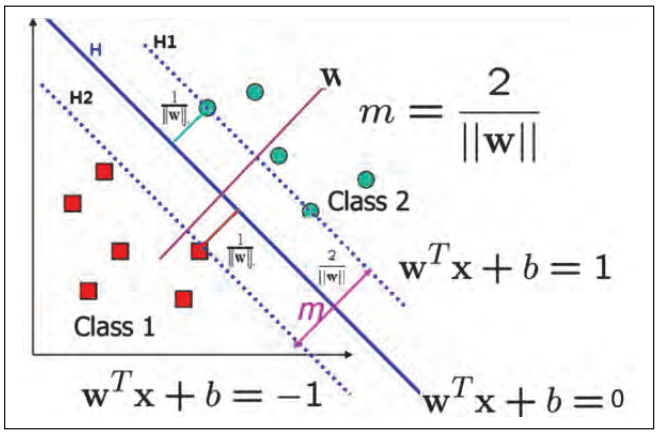

Fig. 2: Schematic diagram of hyperplane in SVM

The sample points on the two hyperplanes are also the points closest to the separating hyperplane in theory. Their existence determines the positions of H_1 and H_2, and supports the dividing lines. They are the so-called support vectors. This is support. The origin of vector machines.With these two hyperplanes, you can logically define the margins mentioned above. In the two-dimensional case, the distance between the two parallel lines ax + by = c_1 and ax + by = c is:

$$\frac{|c_1 - c_2|}{\sqrt{a^2 + b^2}} \tag{4}$$

It can be deduced that the interval between two hyperplanes of H1 and H2 is 2/||w||, that is, the purpose is to maximize this interval.

Assuming that the hyperplane can correctly classify samples, we can make:

$$\begin{cases} w^t x_i + b \geq +1,\ y_i = +1 \\ w^t x_i + b \leq -1,\ y_i = -1 \end{cases} \tag{5}$$

So support vector machine is also called maximum margin hyper plane classifier equivalent to minimizing ||w||, for later derivation and calculation convenience, it is further equivalent to minimization:

$$\frac{1}{2}\|w\|^2 \tag{6}$$

The two formulas are:

$$y_i(w^t x_i + b \geq 1) \tag{7}$$

This is the constraint of the objective function. Now this question becomes an optimization problem:

$$\begin{cases} \min \frac{1}{2}\|w\|^2 \\ subject\ to\ y_i[(w^t x_i + b)] - 1 \geq 0 \end{cases} \tag{8}$$

We can apply this PSO algorithm to multi-objective optimization of this formula. For the above optimization problem, we first need to construct a Lagrangian function:

$$\xi(w, b, a) \equiv \frac{1}{2} w^t w - \sum_{i=1}^{N} a_i y_i (w.x_i - b) \tag{9}$$

Finding w and b separately yields:

$$w = \sum_{i=1}^{N} a_i y_i x_i, \sum_{i=1}^{N} a_i y_i = 0 \tag{10}$$

Then the dual problem of the original problem is obtained by substituting the Lagrange function:

$$\max W(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i=1, j=1}^{n} a_i a_j y_i y_j x_i^T x_j$$

$$subject\ to\ a_i \geq 0, \sum_{i=1}^{N} a_i y_i = 0 \tag{11}$$

3. Water big data classification algorithm based on PSO-SVM hybrid optimization

In this paper, we use heuristic algorithm PSO to optimize the parameters of SVM algorithm, and use grid search to find the best parameters c and g. Although the grid search can find the highest classification accuracy in the sense of CV, that is, the global optimal solution, sometimes it may be time consuming to find the best parameters c and g in a larger range. The algorithm can find the global optimal solution without having to traverse all the parameter points in the grid. The specific process of SVM generation algorithm based on multi-objective PSO clustering is as follows:

Step 1: Take all the training sample sets as the initial root node, call self-mutation PSO clustering algorithm at the root node, and divide the original training samples into two classes to form two child nodes.

Step 2: Determine if the child node contains a class. If it is to Step 4, go to Step 3 if it is not.

Step 3: Continue to call the self-mutation PSO clustering algorithm for this sub-node, subdivide it into two sub-nodes, and go to Step 2.

Step 4: This node is a leaf node and the algorithm ends.

The specific workflow of SVM parameter optimization method based on PSO is shown in Fig. 3.

This method is to determine the corresponding position and training sample of each sub-classifier for each sub-node before the SVM training. The introduction of PSO clustering algorithm will prolong the training time of SVM, but in practical application, the structure optimization and training process of the decision tree is often an off-line process. It is worth and reasonable to change the time consuming of the off-line process in exchange for the better SVM classification performance.

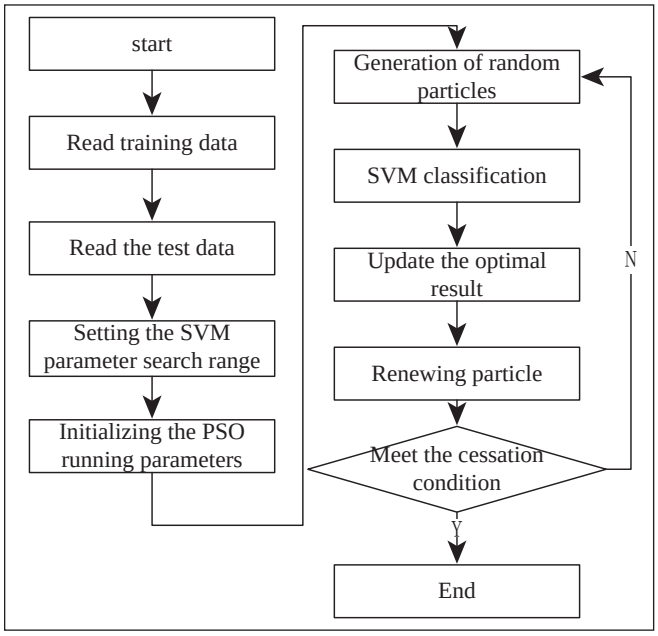

Fig. 3: Schematic diagram of hyperplane in SVM

4. Experimental analysis

This experiment uses the data stream of Yulin water resources big data as the experimental data. According to the SVM classification principle mentioned above, PSO operation steps and improved particle update algorithm, the termination condition is not improved in the 50 generation.

In each experiment, 6000 points were selected as training samples and 6000 as samples to be identified. Considering the randomness of the PSO algorithm, the optimal value is taken as the final result after 10 operations. The optimal result is cross validation. Using the model of 2012 to December 2017 is selected as training samples for monthly water resources level is simulated, and the test from 2012-2016, the monitor water level, the results of the contrastive experiment are shown in Table 1. Among them, the SVM optimal parameter is the result of using the own optimization tool. The search results are shown in Fig. 4.

As can be seen from Table 1, from the perspective of classification accuracy, the proposed data stream classification model has the best classification accuracy in different sub-data and exceeds other single-model classifiers. This shows that compared to other single-model structures with the integrated learning model, the proposed method is effective for the classification of data stream massive data, and can improve the classification accuracy of massive data. From the perspective of time, the proposed SVM-PSO model is not the fastest. This is because ensemble learning uses multiple classifiers, so the classification time will increase accordingly. However, compared to the other two ensemble learning methods, the proposed method is less time consuming than its categorization, and therefore it has improved in time efficiency.

Table 1. Simulation results of bp-pso model

Classification and parameter optimization method	c=200, g=0.6	SVM optimal parameters c=2.225e-2, g=0.205	PSO optimal parameters c=0.002000, g=4.950052
Cross-validation(%)	80.45	81.20	81.32
Number of iterations	8235	1243	656
SV number	1231	2123	4120
Training time (s)	0.5512	0.9861	0.1180
Optimization time (s)	0.4331	0.3456	0.1345
Classification rate (%)	78.21	78.97	79.92

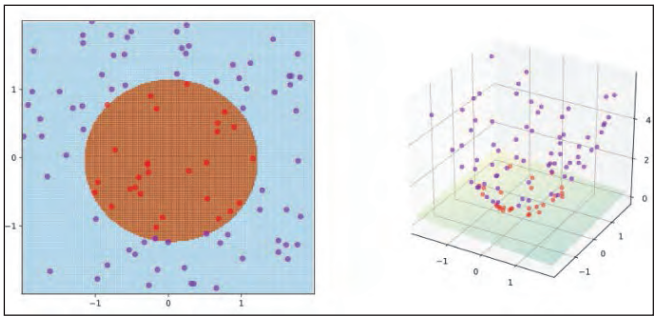

Fig. 4: A training example of SVM with kernel given by PSO

5. Conclusions

In this paper, an improved SVM parameter optimization algorithm based on PSO is proposed, that is, the adaptive convergence speed factor is used to search quickly in the early stage, and the advanced search is carried out in the late stage, thus the requirements of diversification and centralization are met. The SVM multi classifier is used to classify different data. The experimental results show that this method improves the classification accuracy and reduces the classification time to a certain extent, and has certain practical value. Because of the introduction of particle swarm optimization, the relative training time will be extended, but in practical application, the optimization and training process of SVM is often an off-line process. It is worth and reasonable to change the time consuming of the off-line process to get a better SVM classification performance.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 11641002, 51651901); Natural Science Basic Research Plan in Shaanxi Province of China (2015SF261, 2017NY-134, 2016NY141, 2016KJXX-62), Funding Project for Department of Yulin University (16GK24,13YK50), Natural Science and Technology Project Plan in Yulin of China (2016CXY-12), and thanks for the help.

References

1. Liu S S, Zhang H, Mao Z, et al. (2014): Target detection method based on HRM extracting and SVM, *Foreign Electronic Measurement Technology*, 33(10), 38-41.

(Continued on page 732)