Simulation on potential risk mining model of underlying network dormant data

WangJin[1], ChenWanRu[2]

Chengdu Aeronautic polytechnic, Chengdu Sichuan 610100, China

Keywords: data mining; risk; the underlying network

Abstract. This paper focuses on the potential risk mining method of underlying network dormant data, a potential risk mining method based on bat algorithm optimization BP neural network is put forward. The BP neural network parameter is coded as the individual of bat, and the accuracy of the potential risk mining is viewed as the individual fitness function, then through the simulation of bat flight process to find the optimal parameters of BP neural network, finally according to the optimal parameters to establish the potential risk mining model of underlying network dormant data. The experimental results show that, this algorithm can effectively improve the mining efficiency, and further reduce the error.

Introduction

The traditional potential risk mining method of underlying network dormant data is based on the rule analysis and extraction of the potential risk behavior invasion mode and attack characteristics, so as to establish the corresponding mining rules and patterns, so the adaptability and efficiency of mining of the traditional potential risk mining is poor[1-3].

To this end, a potential risk mining method based on bat algorithm optimization BP neural network is put forward. The BP neural network parameter is coded as the individual of bat, and the accuracy of the potential risk mining is viewed as the individual fitness function [4-6]. Then through the simulation of bat flight process to find the optimal parameters of BP neural network [7-9]. Finally according to the optimal parameters to establish the potential risk mining model of underlying network dormant data [10]. The experimental results show that, this algorithm can effectively improve the mining efficiency, and further reduce the error.

The potential risk mining model of underlying network dormant data based on BA-BPNN

Description of BP neural network. A three layer feedforward neural network can approximate any nonlinear function with arbitrary precision, so the BP neural network only need input layer, hidden layer and output layer. The adjustment formula of network parameters (weights and thresholds) is

$$w_{kj}(t+1) = w_{kj} + \alpha\delta_k H_j \quad (1) \qquad w_{ji}(t+1) = w_{ji} + \alpha\delta_j I_i \quad (2)$$

$$\theta_k(t+1) = \theta_k(t) + \beta\delta_k \quad (3) \qquad \theta_j(t+1) = \theta_j(t) + \beta\sigma_j \quad (4)$$

In the formula, w_{kj} (t+1) and w_{kj} (t) are connection weights of hidden node *j* and output layer node *k* during two following trainings respectively; w_{ji}(t+1) and w_{ji}(t) are connection weights of hidden node *i* and output layer node *j* during two following trainings respectively; *H*j is the output of the hidden layer; node; *I*i is the signal input from the input node i; α and β are learning parameters; θ_k and θ_j are threshold at the output node and hidden nodes *j* respectively; δ_k and σ_j are errors of output layer nodes *k* and hidden layer nodes *j*.

Before BP neural network training is started, the most suitable parameter need to be chosen, bat algorithm (BA) is a new swarm intelligence algorithm, through simulating the biological characteristics of bat in nature, like search and predate prey by ultrasonic, so as to find optimal solutions, which has characteristics of simple model, potential parallel and distributed, therefore, this paper uses BA algorithm to optimize BP neural network parameters, so as to improve the accuracy of the potential risk mining of underlying network dormant data.

Description of bat algorithm. BA is a random search algorithm which simulates the bats use a sonar to detect prey and avoid obstacle in nature, its working principle is: the bat individual of M

population is mapped as m feasible solution in D dimensional problem space, the optimization process and the search are simulated as the individual moving and predating process of bat. The fitness function value of problem is solved to measure the advantages and disadvantages of bat location, the individual survival of the fittest is like the iterative process that better feasible solution replace the poor feasible solution in optimization and search process. The implement process of BA are as follows:

(1) the population initialization, i.e. the bat in a random way to diffuse a set of initial solution in D dimensional space. Specific include: the number of individuals in initial population (the number of bats) m, the maximum pulse volume A0, the maximum pulse rate R0, search pulse frequency range [fmin, fmax], volume attenuation coefficient α, enhancement coefficient of search frequency γ, the maximum number of iterations iter_max.

(2) the position Xi of bats is initialized randomly, and according to the merits of the fitness values to find the optimal solutions x*.

(3) update of the bat search pulse frequency, speed and position. In the process of population evolution, search pulse frequency, speed and position of each individual generation can vary according to the following formula:

$$f_i = f_{\min} + (f_{\max} - f_{\min}) \times \beta \tag{5}$$

$$v_i^t = v_i^{t-1} + (x_i^t - x*) \times f_i \tag{6}$$

$$x_i^t = x_i^{t-1} + v_i^t \tag{7}$$

In the formula, $\beta \in [0, 1]$ is a uniform random number; f_i is the search pulse frequency of bat *i*, $f_i \in [f_{\min}, f_{\max}]$, v_i^t and v_i^{t-1} are the speed of bat at time *t* and *t*-1 respectively; x_i^t and x_i^{t-1} are the location of bat at time *t* and *t*-1 respectively; *x** represents current optimal solution of all bats.

(4) generate uniformly distributed random number rand, if rand>ri, then random disturbance is applied to the current optimal solution to generate a new solution, and cross-border treatment is processed for the new solutions.

(5) generate uniformly distributed random number rand, if rand<Ai and f(xi)<f(x*), the new solution generated in step 4 is accepted, then according to the following formula to update r and Ai:

$$A_i^{t+1} = \alpha A_i^t \tag{8}$$

$$r_i^{t+1} = R_0[1 - \exp(-\gamma t)] \tag{9}$$

(6) fitness values of all bats are sorted, so as to find out the optimal solution and the optimal value.

(7) repeat steps (2) ~ (6), until meet the set optimal conditions.

(8) output global optimal value and optimal solution.

The workflow of BP-BPNN potential risk mining model. The position component of individual bat in BA is regarded as weight value and threshold value of BP neural network, each bat only identify a network, in the network training process, the change of positions of individual bats is the update of weight value and threshold value of corresponding neural network, from the bat position update can search the optimal weights and thresholds of the network, so as to achieve the purpose of network training.

After bats individual is applied to the BP neural network weight value and threshold value, bat's fitness is calculated as follows:

$$f(x_i) = \frac{1}{n_t} \sum_{m=1}^{n_t} (O_{im} - T_{im})^2 \tag{10}$$

In the formula, n_i is the number of training samples; O_{im} and T_{im} respectively represents the actual network output and the desired output of *m* training samples under the network weights and threshold determined by *i-th* bat.

The work steps of the potential risk mining model of underlying network dormant data as follows:

(1) the historical data of the underlying network dormant data being attacked is collected, and carries on the pretreatment.

(2) initialize the network structure: the number of nodes in hidden layer and output layer, input

layer, training samples and testing samples are imported.

(3) initialize the bat group: the number of individuals of bat n, volume A of each bat and pulse frequency r, the position vector x and velocity vector v, the range of bat frequency f, bat position xi, the number of iterations and the error precision.

(4) BP neural network weights and threshold adjustment formula (9) ~ (11) are utilized to update individual best position and the global best position for each bat.

(5) formula (5) ~ (7) are utilized to update the search pulse frequency, velocity and position of bat, updating, so as to obtain the best position and the global best position of the individual.

(6) generate uniformly distributed random number rand, if rand<Ai and f(xi)< f(xi)<f(x*), a new solution generated at step 3 step is accepted, then according to the formula (8) ~ (9) to update ri and Ai.

(7) the fitness formula is used to calculate fitness values f of all bats, and sort them as well to get the fitness value fg of the global best location. If fg achieves network training accuracy (fg<ε) or the current iteration times reach the maximum number of iterations, then iteration is terminated and continue step (8); otherwise, the position of individual extreme value qi and global extreme value qg are calculated for each bat, turn to step (3) to update the velocity and position of the bat.

(8) output weight value and threshold value corresponding to bats global optimal position, namely optimal initial weights and thresholds of BP neural network.

(9) weights and thresholds corresponding to the global optimal position of the bat are regarded as BP neural network parameters to build the optimal potential risk mining model of underlying network dormant data.

Simulation experiments

In order to verify the validity of this algorithm, there is the need for an experiment. The experimental data comes from KDD CUP 99 data sets, used for potential risk mining test data sets of underlying network dormant data, where all the data is collected from the actual network data in the Internet, one of it is selected and recorded as follow:

O, tcp，http，SF,1 8 1，5450，0，0，0，0，0，1，0，0，0，0，0，0，0，0，0，0，8，8，0. 00，0. 00，0. 00，0. 00，1. 00，0. 00，0. 00，9，9，1. 00，0. 00，0. 1 1 20. 00，0. 00，0. 00，0. 00，0. 00，normal.

Each record consists of 42 attributes, the last attribute identifier shows whether it is normal data or potentially dangerous data.

With the traditional algorithm and the algorithm for potential risk mining of underlying network dormant data, the error of mining can be described with the graph below:

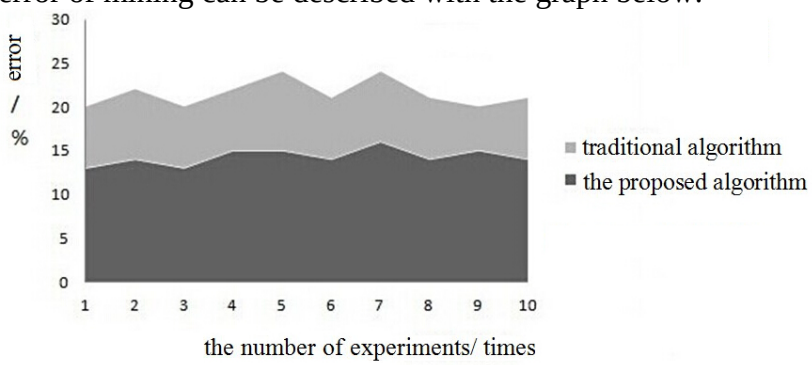

Fig. 1 comparison diagram of mining error of different algorithm

In the above experimental process, the time of mining can be described by the following figure:

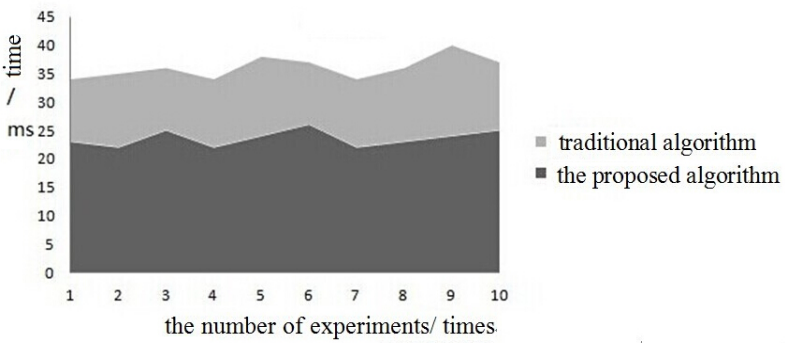

Fig. 2 comparison graph of mining time obtained with different algorithms

The above experimental data were collected and analyzed, the following table is able to be obtained:

Table 1 potential risk mining data table of different algorithms

the number of experiments	traditional algorithm		the proposed algorithm	
	error（%）	time（ms）	error（%）	time（ms）
1	13	23	7	11
2	14	22	8	13
3	13	25	7	11
4	15	22	7	12
5	15	24	9	14
6	14	26	7	11
7	16	22	8	12
8	14	23	7	13
9	15	24	5	16
10	14	25	7	12

Based on the data in the table, it can know with the proposed algorithm for potential risk mining of underlying network dormant data can reduce the error of mining, shorten the time of mining, and further ensure the security of underlying network dormant data.

In order to further verify the effectiveness of the algorithm in this paper, using different algorithms to potentially dangerous mining, actual danger and the relationship between the mining results can be used in the table below, according to the following table could learn, using the results of the algorithm in this paper, mining and the actual danger even closer.

Table 2 relationships between different algorithms and the actual danger data tables

Test times/time	The actual danger/time	Traditional algorithm mining results	The improved algorithm mining results
1	12	6	11
2	14	8	12
3	8	4	6
4	11	7	9
5	16	11	14
6	12	8	11
7	14	9	13
8	12	8	9
9	16	12	15
10	12	6	10

Conclusions

Considering the shortcomings of the traditional potential risk mining method of of underlying network dormant data, a potential risk mining method based on bat algorithm optimization BP neural network is put forward. The BP neural network parameter is coded as the individual of bat, and the accuracy of the potential risk mining is viewed as the individual fitness function, then through the simulation of bat flight process to find the optimal parameters of BP neural network, finally according to the optimal parameters to establish the potential risk mining model of underlying network dormant data. The experimental results show that, this algorithm can effectively improve the mining efficiency, and further reduce the error.

References

[1] Guo Hongyan, Gu Baoping. The application research of Improved K average algorithm in network intrusion detection [J]. Computer security, 2008, 5:24-26.

[2] Liu Jing. The research of network intrusion detection based on clustering [D]. Taiyuan: Taiyuan University of Technology, 2008:20-23.

[3] Qin Ziyan. The research of intrusion detection method based on clustering analysis [D]. Wuxi: Jiangnan University, 2008:42-44.

[4] Liu Chunping. Comparison of the methods based on Kohonen neural network clustering in remote sensing classification [J]. Computer simulation, 2006, (7): 1744-1746.

[5) Wu Ke, Fang Qiang, Zhang Junling, et al. The classification of remote sensing image based on improved Kohonen neural network [J]. Journal of Geomatics, 2007, 32 (2): 47-49.

[6) Li Hui, et al. Network intrusion detection based on support vector machine. [J]. Journal of computer research and development, 2003, 40 (6):799-807.

[7] Chen Xiaohui. Intrusion detection method based on data mining algorithm [J]. Computer Engineering, 2010, 36 (17): 72-73.

[8] Zheng Zhijun, Lin Xiaguang, Zheng Shouqi. A method for data mining based on neural network [J]. Journal of Xi'an University of Architecture and Technology, 2000, 32 (1):28-30.

[9] Li Haiyan, Peng Shimi. Study on the application of genetic neural network in low permeability reservoir diagenetic reservoir facies [J]. Oil and gas geology, 2006, 27 (1):111-117

[10] Wang Anhui, Yu Shuying, Zhang Yingkui, et al. Application of neural network in well test and interpretation in low permeability oil field [J]. Oil and gas geology 2004, 25 (3):338-343.