Network intrusion detection model based on genetic ant colony algorithm

Jianghao huang

Hainan college of software technology, hainan, qionghai, 571400

Keywords: intrusion characteristics; genetic algorithm; ant colony algorithm; pheromone concentration

Abstract. The traditional network intrusion detection is performed on single-dimensional data feature of invasion, once the intrusion has intrusion feature of abnormally high-dimensional data, which can not achieve a unified detection rules, resulting in decreasing efficiency and accuracy of detection. This paper proposes a network intrusion detection method based on genetic ant colony optimization algorithm. According to genetic algorithm building individual coding, employing fitness function to initialize the population, setting pheromone of ants and establishing global pheromone updating rules by ant colony state transition rules, and then ultimately intrusion detection network is accomplished. Experimental results show that modified algorithm for network intrusion detection can improve the speed of training and testing, with significant advantages on increasing detection rate and reducing fault rate.

Introduction

With the rapid development of information technology, network sharing and openness widening, more and more businesses and individuals dependent on the network, hence arising invasion problems of network security is under increasing threat, which has extremely adverse impact on enterprises and individuals' interest, therefore, effective network intrusion detection method has become focused subject by experts and scholars [1-3]. Regular network intrusion detection methods are mainly based on artificial immune algorithm [4], neural network algorithm [5] and rough set algorithm. Effective network intrusion detection method can improve the efficiency and detection rate, ensure network security, so as to become hot issue in this intrusion field with broad prospects.

Network intrusion detection principles

For high-dimensional abnormal invasion feature complete network intrusion detection is of great significance. Specific principles are described as follows:

Network intrusion detection is based on machine learning theory, building normal and abnormal sample library, through machine learning of normal behavior and abnormal behavior patterns to judge effectively.

Assumed x as the independent variable, y as the dependent variable, given that x and y has unknown relationship, their joint probability distribution $F(x,y)$ is unknown, $(x_1,y_1),(x_2,y_2),(x_3,y_3),...,(x_n,y_n)$ describes observed sample of n independent distribution, risk function of training process is:

$$R(d)=\int S(y,f(x,d))dF(x,y) \quad (1)$$

As $f(x,d_0)$ represents the minimum optimal function， and satisfies $(f(x,d_0)\in\{f(x,d)\})$. $\{f(x,d)\}$ which describes the set of candidate prediction function, as d is function parameters. $S((y,f(x,d))$ describes loss function, indicating difference between actual output $f(x,d)$ and the desired output with fixed x .

Due to unpredictability of $F(x,y)$, sample empirical risk is employed as approximate expectation risk:

$$R_{emp}(d)=\frac{1}{n}\sum_{i=1}^{n}S(y_i,f(x_i,d)) \quad (2)$$

Considering training classification, function set is specified, the relationship between experiential risk and actual risk can be expressed as

$$R(d) \le R_{emp}(d) + \sqrt{\frac{h(1n(2n/h)+1-1n(A/4))}{n}} \quad (3)$$

The formula can be simplified to:

$$R(d) \le R_{emp}(d) + \phi(\frac{n}{h}) \quad (4)$$

By minimizing the summation of two terms on the right side of formula, adjusting VC dimension, reducing the value h , minimizing risk, improving generalization ability training, the accuracy of network intrusion detection is approved.

GA-ACO Task Scheduling Algorithm

Task scheduling design of genetic algorithm.

1)Individual coding

Chromosome coding is the key to solving CMP task scheduling. In order to take full account of the interdependence of relevant factors and the location information of each sub-task, a three-dimensional coding scheme is proposed as follows:

$$ST = \begin{Bmatrix} List_{11} & List_{12} & \cdots List_{1p} \\ List_{21} & List_{22} & \cdots List_{2p} \\ \cdots & \cdots & \cdots \cdots \\ List_{n1} & List_{n2} & \cdots List_{np} \end{Bmatrix} \quad (5)$$

Where, Listij means subtasks sequences executing on the processor.

According to equation (4), the GA's population is individually encoding, which provide a good description of CMP task scheduling program.

2) Fitness function

CMP scheduling takes task completion time span (Makespan) into major consideration, fitness calculated by fitness function should be non-negative and generally large value for more excellent individuals [9].

Therefore, maximum makespan value is selected from the population of individuals 2*Tmax minus Makespan[i] value as a fitness function, so the fitness function Fit(i) of individual i is defined as

$$Fit(i) = 2 * T_{\max} - Makespan[i] \quad (6)$$

Where, Tmax is largest makespan value in individual population.

3) Population initialization

Population initialization is the key to GA, outstanding population saves computational time; however traditional random algorithm induces locally optimal solution, proposed method improves in initial producing way, as follows:

Assuming randomly generating N individual, with length K, and then to compare similarity of two individuals, similarity is defined as:

$$Similar(i, j) < \frac{Num_{same}(ij)}{K} \quad (7)$$

It is required that similarity between individuals should meet following conditions,

$$Similar(i, j) < \frac{K-c}{K} \quad i \ne j \quad (8)$$

Where,c is adjusting constant to control expectation length.

With the generation of initial population can not only ensure that all particles have obvious divergence, but also evenly distributed, which effectively reduce the probability of local optimum and improve the chance of finding the global optimal solution.

4) Genetic Operators

According to the pre-set probability optimum individuals are chosen as the next generation, other individuals as crossover for mutation operation to make new individual by single-point crossover and random mutation.

Ant colony algorithm design.

1)Ant pheromone setting

CMP scheduling option obtaining from GA is converted to ACO initial pheromone, in particular:

$$\tau_{ij}(\alpha) = fit(x_i) \tag{9}$$

2)ACO state transition rule

After a period ACO need to select a crawling node on the basis of pheromone concentration, at time t, the state transition rule of the ant colony i is

$$p_{ij}(t) = \begin{cases} \dfrac{[\tau_{ij}(t)]^{\alpha}}{\sum_{j \in allowed} [\tau_{is}(t)]^{\alpha}}, & rand(0,1) \le \rho \\ Rand(allowed), & rand(0,1) > \rho \end{cases} \tag{10}$$

Where,ρ represents a random number between 0 and 1, with the expression as

$$\rho = 0.1 \times \frac{\log(9) \times t)}{e^{t_{max}}} \tag{11}$$

Where, t and t_{max} respectively presents current and maximum value.

3)Pheromone global updating rule

$$\Delta\tau_{ij}(t,t+1) = \begin{cases} Q, \ if \ \text{ant k through ij} \\ 0 \ , otherwise \end{cases} \tag{12}$$

CMP scheduling steps. CMP scheduling steps based on GA-ACO algorithm are as follows:

Step1： Set GA parameter, such as population size, the maximum evolution generation, mutation probability and algorithms end condition;

Step2： Population initialization and coding;

Step3： Calculate individual fitness value based on fitness function;

Step4： Generate new individuals by selection, crossover and mutation operations;

Step5： If GA end condition is satisfied, then skip to Step6, otherwise to Step3;

Step6： Obtain initial value of ACO pheromone converted by optimal solution of CMP scheduling;

Step7： The ant is placed on the graph nodes;

Step8： According to the state transition rules, the next node is chosen for every crawling ant;

Step9： Update every ant's crawling pheromone concentration;

Step10： Calculate value based on the fitness function of ants crawling path, and update its global pheromone concentration;

Step11： If completely conditions of the algorithm are achieved, optimal solution of CMP task scheduling is obtained; if not, skip to Step8 to proceed finding the optimal CMP task scheduling.

Experimental results and analysis

In order to verify the effectiveness of the improved algorithm, it is need to perform an experiment. Experimental procedure for the data source is *kddcup.data_10_percent* data set of kddcup1999.Experiment data are randomly selected as the training sample set, as shown in Table 1, the training data classified as I_1、I_2、I_3、I_4 respectively, using the traditional algorithm and improved algorithm for network intrusion detection simulation.

Table 1 Experimental training data

Attack type	Sample numbers	Regular sample	Attack sample
Probe	6700	5200	1500
U2L	4800	3987	813
U2R	2760	2320	440
DoS	1755	1690	65

Comparison between improved and traditional algorithms on detection rate is shown in Figure 2:

It is demonstrated in figure1 that with small amount of detecting data, detection rate between the traditional algorithm and improved algorithm difference is small, while with the increase of test data significant difference occurs and improved algorithm has significant advantages.

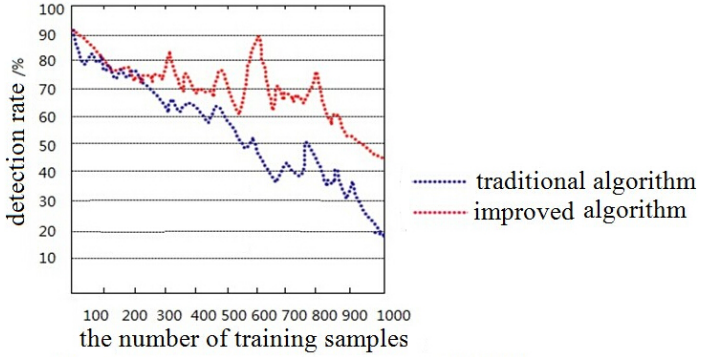

Figure1 Comparison between different algorithms

Statistics results conducted on the traditional algorithm and improved algorithm are as follows in Figure 2, which indicates that improved algorithm has better performance and efficiency than traditional one.

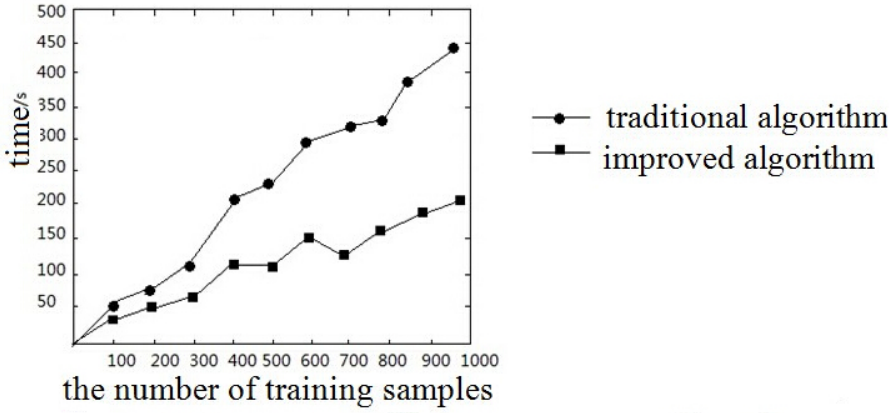

Figure 2 Comparision of training time

To investigate performance of the impact factor η, two types of attacks DoS, U2R are selected for experiments. Detection rate and time are demonstrated in Figure 3, Figure 4 respectively:

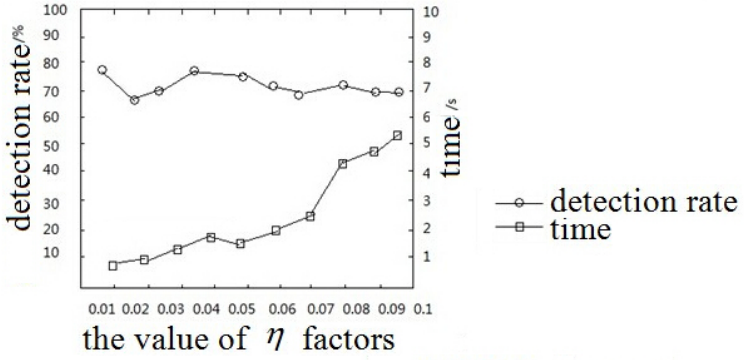

Figure3 Algorithm performance by factor η for DoS attack type

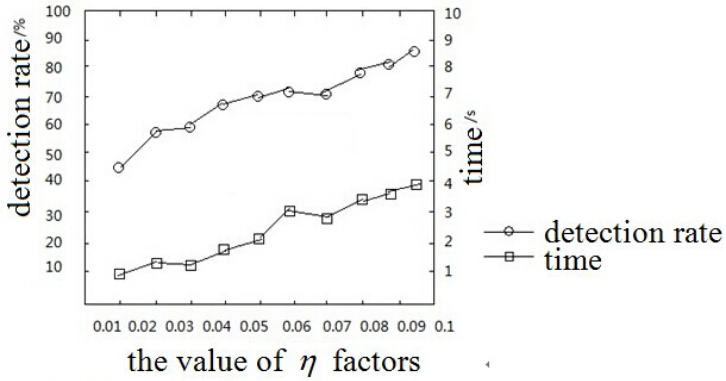

Figure4 Algorithm performance by factor η for U2R attack type

From Figure 3, Figure 4 it is seen that in the type of DoS attack, when values ranges from 0.01 to 0.05, detection efficiency is higher and expense less time-consuming, to achieve the best performance of the algorithm. In the type of U2R attack when values range from 0.05 to 0.07, the best performance of the algorithm achieves. It is indicated that during intrusion detection, setting proper factor η algorithm can win effectively and reduce training time to achieve the best performance.

Conclusions

Traditional network intrusion detection is mainly for single dimensional data feature, once the invasion with abnormally high dimensional data feature can not be formed by a unified association detection rules, resulting in inefficient and inaccurate detection rate. This paper proposes a network intrusion detection method based on genetic ant colony optimization algorithm. Genetic algorithm based on the individual components coding, using the fitness function to initialize the population, setting pheromone of ants and ant establishing colony state transition rules to update global pheromone rules, and ultimately intrusion detection network is performed. Experimental results show that the improved algorithm for network intrusion detection has significant advantages, not only can improve the detection rate also can effectively reduce training time, to protect network security and meet the actual demand.

References

[1] Pan Julong, Li Shanping, Zhang Daoyuan. Game detection method for cluster suspicious nodes in wireless sensor network [J]. Journal of Zhejiang University: Engineering Science, 2012,46(1):72-78.

[2] TANG Rui,ZHAO Jihong,QU Hua.Distributed Power Control for Energy Conservation in Hybrid Cellular Netwok with Device-to-Device Communication[J]. China Communications, 2014(3):27-39.

[3] Chen Hongxing. The characteristics of network intrusion detection based on genetic algorithm optimizing neural network [J]. Computer engineering and applications, 2014(14):78-81.

[4] Xu Zhenhua. The characteristics of network intrusion detection based on sparse fuzzy positioning algorithm [J]. Bulletin of science and technology, 2014, 30(2):233-235.

[5] Mitra Montazeri,Hossein Nezamabadi-pour, Mahdieh Montazeri.Automatically Eye Detection with Different Gray Intensity Image Conditions[J].Computer Technology and Application, 2012,3(8):525-532.