Feature Selection Algorithm Based on Quantum Evolution in Network Intrusion Detection

[1]Haogui CHEN, [2]Huifeng LONG
[1]Modern Education Technology Center, Hunan City University, HuNan, Yiyang 413000, CHINA
[2]College of Urban Management, Hunan City University, HuNan, Yiyang 413000, CHINA

Abstract: As there are ubiquity problems about slowing in detecting limited by optimizing performances in current network intrusion detection, this paper proposed the feature selection algorithm based on quantum evolution. Firstly, in order to reach optimizing performance, this algorithm promoted quantum evolution algorithm for objective and formed the evaluation function of feature subset, then designed network intrusion detection feature selection algorithm according to the process of quantum evolution algorithm, finally made a comparing experiment of the algorithm in this text with feature selection algorithm based on genetic algorithm(FS-IGA), the results showed: the global optimizing ability of quantum evolution's feature selection algorithm made universal search to feature space, took out unrelated and useless features, increased detection effect.

Keywords: Intrusion Detection; Optimized Feature; Coding, Detection Rate

1. Introduction

Intrusion detection system (IDS) is an significant part of deep network defense system, according to detecting and analyzing related audit data like network flow, system log and so on, the IDS judges whether there are actions like violating security policy or computer's system security [1-3]. Intrusion detection is to detect intrusion behavior, through collecting information and key points from the computer network or computer system analysis, from which to discover whether there is a breach of security strategy and plans to attack the network or system. Its main function is to monitor the running state of the system, find out all kinds of attacks, aggression or attack results, to ensure that the resources of the system confidentiality, integrity and availability [4]. Because intrusion detection can monitor the network in case of not affecting the network performance, provide effective protection for internal attack, external attacks and misuse of the system. Therefore, to provide efficient intrusion detection and the corresponding protective measures for network security, such as recording evidence for tracking and recovery, disconnect the network connection. It with the control detection for the essence of technology, can make up for the shortage of firewall, plays a role of active defense, is an important part of network security. Intrusion detection is a new security mechanism begins to integrate into the security framework of network system. Feature subset algorithm can choose about 10 d, compared with the original feature set, using the feature subset in the intrusion detection model is established and the efficiency of intrusion detection has significantly improved, and the algorithm for feature selection efficiency is superior to the literatures of the intrusion detection feature selection based on genetic algorithm, the algorithm. Anderson is the one who firstly proposed the concept of intrusion detection, Denning and his colleagues proposed a more complete model based on statistics, Mark off chain and time series. Many learning algorithms were applied into intrusion detection and those algorithms can be divided into incremental algorithm and non-incremental algorithm [5-7].

With the expanding and popularizing of the network, network attack has becoming more and more, these attacking can lead to unavailable of network even worse result, network security now is a global-focused problem. Network intrusion detection is a new generated network security technology based on positive defense strategy, it detects whether there are intrusion actions according to analyzing feature characteristics in network connecting, and this is a very significant technology in current network security fields [8-10]. However, with the increasing speed of network, the prominent issue faced by network intrusion detection is the mass data package cannot be real-time processed in network connecting. Related researches showed, too much of network connecting feature information of selecting and analyzing in network intrusion detection is the main reason of lower speed of detection [11-14]. For this, some scholars worked on decreasing feature dimension of network connecting by feature selecting and to solve the problems of low detection speed in network intrusion detection, which made good results. Network intrusion detection is a new generated network security technology based on positive defense strategy and it’s also the most important link in network security structures, it has being wildly used in detecting, recognizing and tracking the intruders. Intrusion detection, as a positive and real-time security protection technology, it not only makes up the disadvantages of firewall, data encryption, certification and other static defense strategies, but also offers all-round protection for network with other network security strategies [15-16].

In recent years, domestic researchers has turned the feature selecting of network intrusion detection into optimizing, used optimized algorithm to get feature subset and then used it in intrusion detection, as Intrusion Detection's Feature Selection Based on Genetic Algorithm respectively used genetic algorithm and simulated annealing algorithm to select the feature subset; A Quick Feature Selection Method of Intrusion Detection used particle swarm algorithm to get the feature subset, compared with original feature set, it can increase the effect of intrusion detection. While the complexity of network connections data structure and its high dimension, it’s an NP combinatorial optimization problem of selecting effect feature subset from the original feature set of network connecting, and it leads to the optimizing performance of those optimizing algorithm limited and hard to get effect feature subset, so the integrated detection performance needs to be promoted.

Quantum evolutionary algorithm (QEA) is the product of quantum computing theory and evolutionary computing theory, a now-developed intelligent optimization algorithm. The special coding and evolution mechanism of QEA gave it a superiority performance in solving combinatorial optimization problem. This text uses QEA to select the feature characteristics in network connecting, uses the global optimizing ability of QEA to make global search for feature space and to take out unrelated and useless feature characteristics, gets optimized feature subset and uses in intrusion detection to increase the detection rate.

This paper mainly made expand and innovative work in the following areas to:

(a) The network connection has complex data instruction and high feature dimension, selects effect feature subset from original feature set of network connecting is an NP combinatorial optimization problem of selecting effect feature subset from the original feature set of network connecting, and it leads to the optimizing performance of those optimizing algorithm limited and hard to get effect feature subset, so the integrated detection performance needs to be promoted. This text proposed feature selecting algorithm based on quantum evolution in coder to solve this problem. This algorithm uses quantum evolutionary algorithm in the feature selecting of network intrusion detecting, selects a set of efficient feature from the original feature characteristics of network connecting and applies to intrusion detection in order to increase detection rate. Firstly promotes quantum evolution algorithm by aiming at increasing optimization ability, Fisher rate based on feature characteristics forms the evaluation function of feature subset, then designs the feature selection algorithm of network intrusion detection by quantum evolutionary algorithm. In The algorithm can select a feature subset of 10 dimensional or so, compared with the original feature set, use the feature subset has been significantly improved in efficiency by building models of intrusion detection and intrusion detection, and feature selection algorithm efficiency is better than that of "genetic algorithms to intrusion detection algorithm based on feature selection".

(b) In order to further prove the accuracy and efficiency of feature selecting algorithm based on quantum evolution, makes a comparison experiment between the algorithm in this text with feature selection algorithm based on genetic algorithm(FS-IGA), the laboratory uses 41 dimension characteristic which takes TrD1 as training sample and sample instances as original feature set, makes feature selection by the algorithm designed in this text, Compared intrusion detection model based on 9 dimensional feature subsets with original intrusion detection model based on 41 dimensional features, the modeling time and test time are significantly reduced invasion. In the detection rate and false alarm rate, the detection rate of the former increased an average of 0.996% than the latter; in the feature selection algorithm spending time, FS-IQEA is less 79%than FS-IGA. Compared the detection model based on 10 dimensional feature subsets with original intrusion detection model based on 13 dimensional features, the modeling time and intrusion test time overhead are slightly less. The experiment result showed: in the feature selection of network intrusion detection, the optimization ability and speed of the algorithm in this text is better and faster than the feature selection algorithm based on genetic algorithm(FS-IGA), while takes out unrelated and useless feature characteristic, increases detecting effect.

2. Feature Selecting

The mathematic model of feature selecting: for an given feature set F = { f1,f2,…,fn} , n is the range of feature set, one of its feature subset can be expressed by a binary vector: S = { s1,s2,…, s_n } ($s_i \in \{0,1\}$, $i=1,2,...,n$) , s_i = 1 refers to i feature f_i is selected; contrary, it refers to i feature f_i is not selected. Feature selecting needs to solve the two key problems: designs searching strategy and evaluating function, whose effect is generates and evaluates feature subset.

2.1. Searching Strategies

Nowadays the basic searching strategies of feature selecting is exhaustive search, random search and heuristic search, they both have advantages and disadvantages, and always uses in combination. QEA, as a randomly probability searching algorithm, because it takes heuristic search strategy in computing process, which makes its searching space is far less than O(2n), and has special advantages in feature selecting, but its optimize performance is not ideal when solving complex problem, so QEA needs to be promoted and solving problems using promoted QEA.

Population evolution is the driven mechanism of QEA optimizing, often uses quantum revolving door of formula (1) and to update quantum bit through formula (2) to realize evolutionary operations, the value of q_i in updating process is got by previously defined rotation angle table, the angle in this table is discrete and settled, the updating operates for chromosome's quantum bit is limited in several settled conditions.

$$t(q)=\begin{bmatrix}\cos(q)-\sin(q)\\ \sin(q)\cos(q)\end{bmatrix} \qquad (1)$$

There, θ is rotation angle.

$$\begin{bmatrix}e_i^{i+1}\\ v_i^{t+1}\end{bmatrix}=t(q_i)\begin{bmatrix}e_i^{j}\\ v_j^{i}\end{bmatrix}=\begin{bmatrix}\cos(q)\\ \sin(q)\end{bmatrix}\begin{bmatrix}e_i^{j}\\ v_j^{i}\end{bmatrix} \qquad (2)$$

$[e_i^i, v_j^i]$ T is the probability amplitude uses to update i quantum bit of the chromosome; $[e_i^{j+1}, v_j^{i+1}]$ T is the probability amplitude of i quantum bit of the chromosome after updating; $q_i = s(e_i, v_j).\Delta q_i.s(e_i, v_j)$ respectively refers to the dimension and direction of rotation angle.

In complex optimize issue, this condition limited updating operations has limited optimizing ability and cannot search for global optimization, while it needs complex multipath conditions for judging in searching q_i, this affects the efficiency of algorithm. So the promotion of computing for q_i is the key to increasing the optimizing ability of QEA and also the main direction of promoting QEA. The promoted strategies' design process in this text is as following:

Proposes angular distance of qubit phase angle.

For combinatorial optimization, often simplifies the probability of quantum bit as real number, so there uses a set of real number (α,β) which meets the conditions of normalization to express the probability amplitude of one quantum bit, now the quantum bit can reflect as unit vector in 2-D real space, and the quantum bit in quadrant two, three and four can project in quadrant one, and gives the definition of quantum bit's phase angle and angle distance.

Definition one: Supposes the probability amplitude of given quantum bit $|J|$ is $|a, b|$ T (a, b is real number), supposes $\arctan(|b|/|a|)$ is the phase angle of $|J|$, sets $v\ |J| \in [0, p\ /\ 2]$.

$$v\ |J|=\begin{cases}\arctan(|b|/|a|), |a|\neq 0\\ p\ /\ 2, |a|=0\end{cases} \qquad (3)$$

From definition one, the phase angle of $|0|$ is 0, the phase angle of $\langle 1\rangle$ is $p\ /\ 2, ie, v\ |J|=0, v\ \langle 1\rangle = p\ /\ 2$ π/2.

Definition two: Supposes phase angle of quantum bit $\langle J\ 1\rangle$ is $v\ \langle J\ 1\rangle$, phase angle of $|J\ 2\rangle$ is $v\ |J\ 2\rangle$, calls $v\ |J\ 2\rangle - v\ |J\ 1\rangle$ is the angle distance from $|J\ 1|$ to $|J\ 2\rangle$, sets:

ΔJ

$|J1\rangle \rightarrow |J2\rangle$

Obviously angle distance is a vector, including magnitude and symbols, magnitude refers to the range of deflection, symbols refers to the direction of deflection, "+" refers to clockwise, "-" refers to anticlockwise. From definition two, two special angle distance as the formula (4) and (5):

$$\underset{|J\rangle\rightarrow|0\rangle}{\Delta J} = v_{|0\rangle} - v_{|j\rangle} = -\arctan(|b|/|a|) \qquad (4)$$

$$\underset{|J\rangle\rightarrow|0\rangle}{\Delta J} = v_{|1\rangle} - v_{|j\rangle} = p\ /\ 2 - \arctan(|b|/|a|) \qquad (5)$$

Using angle distance calculating rotation angle.

Analyzing the quantum bit's updating process of formula (2), under the effect of quantum revolving door, probability amplitude of quantum bit changes, and changes the probability of values 1 or 0 of quantum bit, in reality, it makes the quantum bit deflect, and proposes a dynamic adjust q_i strategy based on angle distance. The rotation angle can be calculated by formula (6) and (7).

$$q_i = (1 - f_x\ /\ f_b)\underset{|J_i\rangle\rightarrow}{\Delta J} \qquad (6)$$

There: f_x is the fitness value of updated individual, f_b is the fitness value of current optimal individual.

$$\underset{|J_i\rangle}{\Delta q} = \begin{cases}\underset{|J_i\rangle\rightarrow|0\rangle}{\Delta q} = -\arctan(|b_i|/|a_i|)\\ \underset{|J_i\rangle\rightarrow|1\rangle}{\Delta q} = p\ /\ 2 - \arctan(|b_i|/|a_i|)\\ (f_b \geq f_i)\wedge(b_x = 1)\wedge(x_i = 0)\\ 0, others\end{cases} \qquad (7)$$

When calculating rotation angle q_i by formula (6) and (7), gets the symbol of $\underset{|J_i\rangle}{\Delta q}$ and can make all the quantum bit of current chromosome deflect to the corresponding quantum bit's ground state of optimal individual, which ensures the evolution direction of the algorithm; while the magnitude of q_i can self-adapted calculates appropriate angle by the condition of current needs to be updated quantum bit and the fitness value of chromosome, then ensures the evaluation performance of the algorithm. So the process of calculating rotation angle q_i by formula (6) and (7) is dynamically and continuously, it's more comprehensive and careful for the searching of the result space, and simple in calculate, easy to understand.

Using H door revising probability amplitude of qubit

Analyzing the adjust strategy in formula (6) and (7) and getting the adjustments still has some limitations, when individuals are far from the optimal individual, $f_b \;/\; f_i$ is very small, and it needs to be adjusted a lot, probably makes the probability amplitude of quantum bit fast tends to 0 or 1 and leads to observed value convergences too early, especially in the early times of evolution, most individuals has low fitness degree, too much of adjustment may make them convergences too early and fells into premature convergence. In order to overcome this, using H door to revise probability amplitude of quantum bit after updating, revise strategy is to set a minimum threshold , when $|a|^2$ or $|b|^2$ of quantum bit after updating is less than threshold ε, uses formula (8) to revise this quantum bit.

$$\left[a_i^{l+1}, b_i^{l+1}\right] = \begin{cases} \left[\sqrt{e}, \sqrt{1-e}^{\,t}, a_i^{l+1} \geq 1-e\right] \\ \left[1-\sqrt{e}, \sqrt{e}\right]^t, \left|b_i^{l+1}\right| \geq 1-e \\ \left[a_i^{l+1}, b_i^{l+1}\right]^t, others \end{cases} \tag{8}$$

Form formula (8), the revising operation of quantum bit is when $|a|^2$ or $|b|^2$ is too much close to 0 or 1, so as to increase the diversity of observing population and avoid premature convergence of the algorithm.

Using chromosomal chiasma to increase the activity of population.

One of QEA'S disadvantages is it sometimes fells into the condition of stop in the last times of evolution which lows the optimize effect of algorithm, using chromosomal cross to promote this, the basic idea is when the algorithm stops, makes small disturbance by cross to the population individuals and changes parts of the chromosome's quantum bit, increases the activity of population, keeps the searching of algorithm going. The judgment of algorithm's stop is: given a relatively small threshold $0 \,\mathbf{p}\, e \quad 1$, if the neighbors two population's average fitness change less than δ in continuously 5 generations, and then regard it as stop, cross should be taken, the steps of chromosome cross is:

Randomly orders the chromosomal in population;

Randomly confirms m(0<m≤5) qubit as chiasma point;

Circular shifts n(number of chromosomal) times of qubit probability amplitude in chiasma point.

2.2. Evaluating Function

Feature selecting can be divided into Filter and Wrapper according to the relationship between feature subset evaluation principles and subsequent classifier. Considering the big data in network intrusion detection and the high dimension in data feature, takes Filter evaluation principle based on distance, uses Fisher rate to design the evaluation function of subset, main idea is the stronger classified feature has shorter cluster distance while weaker classified feature has longer cluster distance.

To the data set $x = \{x_1, x_2, \ldots x_n\}$, $x_i \,(1 = 1, 2, \ldots, n) \in r_d$ makes up by n samples in d-dimension feature space, divides into c class: C1,C2,…,Cc,each has n_i samples.

$$\sum_{i=1}^{r} n_i = n$$

Definition three: the ratio between distribution information among course $s_b^{(k)}$ in sample set in k dimension and distribution information inside a class $s_w^{(k)}$ is called Fisher rate $s_b^{(k)} \;/\; s_w^{(k)}$.

$$s_b^{(k)} = \sum_{i=1}^{r} \frac{n_i}{n} \left(l_i^{(k)} - l^{(k)} \right)^2 \tag{9}$$

$$s_w^{(k)} = \sum_{i=1}^{r} \left(\frac{1}{n_j} \sum_{\substack{x \in r \\ j \sum_{i=1}^{n} x_i^2}} \left(x^{(k)} - m_j^{(k)} \right)^2 \right) \tag{10}$$

There $s^{(k)}$ refers to k dimension feature value of sample x, $s_j^{(k)}$ and $m^{(k)}$ are respectively the average value of k dimension feature of j sample and all samples; $s^{(k)}$ refers to the distances between different kinds of samples, $s_w^{(k)}$ refers to the distances between same kinds of samples.

Definition four: According to definition three, the ratio between distribution information among course $s_b^{(rk)}$ in sample set of the feature subset makes up by k features and distribution information inside a class $s_w^{(rk)}$ is called Fisher rate $s_b^{(rk)} \;/\; s_w^{(rk)}$.

$$s_b^{(rk)} = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{n_j}{n} \left(r_j^{(i)} - r^{(i)} \right)^2 \tag{11}$$

$$s_b^{(rk)} = \sum_{i=1}^{r} \sum_{j=1}^{e} \left(\frac{1}{n_j} \sum_{x \in c_j} \left(x^{(i)} - m_j^{(i)} \right)^2 \right) \tag{12}$$

To simplifying the calculation, there divides two classes of network intrusion detection's data samples: normal data and intrusion data, respectively calls positive samples and negative samples, now the problems is simplified as binary classification problems. To this sample set $x = \{x_1, x_2, \ldots x_n\}$, sets positive samples as 0.0 in X, negative samples as X2, n1 is the number of positive samples, n2 is the number of negative samples, from definition four and gets:

$$s_b^{(rk)} = \sum_{i=1}^{r} \sum_{j=1}^{e} \left(\frac{1}{n_j} \sum_{x \in c_j} \left(x^{(i)} - m_j^{(i)} \right)^2 + \frac{n_2}{n} \left(m_2^{(i)} - m^{(i)} \right)^2 \right) \tag{13}$$

From definition three and four, Fisher rate refers to the feature's contribution degree to classify, a bigger Fisher rate the feature subset has, the stronger classified ability of the feature subset is. So, aims at network intrusion detection, the evaluation function of feature subset can be designed as:

$$S = \sum_{i=1}^{r}\left(\frac{n_1}{n}\left(m_1^{(i)} - m^{(i)}\right)^2 + \frac{n_2}{n}\left(m_2^{(i)} - m^{(i)}\right)^2 \right)$$

$$\sum_{i=1}^{r}\left(\frac{1}{n_1}\sum_{x\in x_1}\left(x^{(i)} - m_1^{(i)}\right)^2 + \frac{1}{n_2}\sum_{x\in x_1}\left(x^{(i)} - m_2^{(i)}\right)^2 \right) \quad (14)$$

CONCLUSION

There are many unrelated and useless feature information in network data package, if they are all used in intrusion detection, the detection efficiency would be decreased, this text used quantum evolutionary algorithm to reduce the feature dimension in network connection, proposed the feature selection algorithm based on promoted quantum evolution algorithm. In the experiment of KDD99 dataset, the algorithm can select about 10 dimension of feature subset, compared with original feature set, it has obviously increase in modeling intrusion detection and its effect using feature subset, and the algorithm's feature selecting effect is better than the algorithm in document Intrusion Detection Feature Selection Based on Genetic Algorithm, increased the detection effect.

References

[1] Kuo-Feng Huang, Shih-Jung Wu, Real-time-service-based Distributed Scheduling Scheme for IEEE 802.16j Networks. Journal of Networks, Vol 8, No 3 (2013), 513-517

[2] Tan x, Triggs B. Enhanced local texture feature sets for face recongnition under difficult illuminationing conditions [J]. IEEE Trans on Image Processing, 2010, 19(6): 1635-1650.

[3] Muhammad J. Mirza, Nadeem Anjum.Association of Moving Objects Across Visual Sensor Networks.Journal of Multimedia, Vol 7, No 1 (2012), 2-8

[4] Haiping Huang, Hao Chen, Ruchuan Wang, Qian Mao, Renyuan Cheng.(t, n) Secret Sharing Scheme Based on Cylinder Model in Wireless Sensor Networks.Journal of Networks, Vol 7, No 7 (20 12), 1009-1016

[5] Muhammad J. Mirza, Nadeem Anjum,Association of Moving Objects across Visual Sensor Networks. Journal of Multimedia, Vol 7, No 1 (2012), 2-8

[6] Yang M,Zhang Lei, "Gabor Feature based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary", Proceedings of the European Conference on Computer Vision, ECCV, pp. 448-461, 2010.

[7] Land E. The Retinex[J]. American Scientist, 1964, 52(1): 247-264.

[8] Kasman Suhairi, Ford Lumban Gaol, The Measurement of Optimization Performance of Managed Service Division with ITIL Framework using Statistical Process Control. Journal of Networks, Vol 8, No 3 (2013), 518-529

[9] Zhao Liangduan, Zhiyong Yuan, Xiangyun Liao, Weixin Si, Jianhui Zhao.3D Tracking and Positioning of Surgical Instruments in Virtual Surgery Simulation. Journal of Multimedia, Vol 6, No 6 (2011), 502-509

[10] Young k P, Seokl P, Joong K K. Retinex method based on adaptive smoothing for illumination base on adaptive smoothing for illumination invariant face recognition[J]. Signal Processing, 2008,88: 1929-1945.

[11] Guang Yan, Zhu Yue-Fei, Gu Chun-Xiang, Fei Jin-long, He Xin-Zheng, A Framework for Automated Security Proof and its Application to OAEP. Journal of Networks, Vol 8, No 3 (2013), 552-558

[12] R. Berangi, S. Saleem, M. Faulkner, et al. TDD cognitive radio femtocell network (CRFN) operation in FDD downlink spectrum. IEEE, 22nd International Symposium on Personal, Indoor and Mobile Radio Communications, 2011: 482-486

[13] Provenzi E Gatta C Fierro M et al. A spatially variant white-patch and gray-world method for color image enhancement driven by local contrast[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(10):1757-1770.

[14] Zhou S, Aggarwal G, Chellapa R. Appearance characterization of linear lamebrain object, Generalized photometric stereo and illumination- Invariant face recognition [J]. IEEE Trans on PAMI, 2007, 29(2): 230-245.

[15] Xin Huang, Xiao Ma, Bangdao Chen, Andrew Markham, Qinghua Wang, Andrew William Roscoe.Human Interactive Secure ID Management in Body Sensor Networks. Journal of Networks, Vol 7, No 9 (2012), 1400-1406