Study on the method of effective extraction of virus feature large joint network

Zhang Zhihua

Educational Technology and Computer Center，Zhaoqing University，Zhaoqing Guangdong 526061,China

Keywords: large joint network; virus feature; effective classification; game factor

Abstract. In large joint network, traditional methods cannot accurately determine the source of virus, leading to data classification of virus feature extraction model in joint network with a low convergence efficiency. A kind of virus feature extraction model for large joint network based on unconstrained clustering correlation and repeated game factor is put forward, according to the identification attribute of access data perfect it. Using unconstrained clustering correlation virus detection algorithm make accurate classification of the multi feature interference in joint network. In the classification probability calculation, constraints computational game factors are introduced. Using data game filter multi-time probabilistic contrast in the features probability matching process of joint network virus. By calculating the optimal reaction function, makes the joint network virus feature extraction to achieve optimal. Simulation results show that, the proposed model can effectively extract the characteristics of joint network virus, and the efficiency and the accuracy is better than the traditional model, has obvious optimization effect.

Introduction

With the rapid development of computer network technology, network type is also increasing, virus invasion of large joint network has the complexity and randomness, greatly reducing the accuracy and efficiency of virus features extraction in joint network [1, 2]. Due to the increase of network type and type of network virus, the traditional methods for network virus feature extraction are unable to real-time to ensure network security, the acquisition of effective classification method for network virus feature extraction has important significance to improve the safety of the large joint network [3].

Large joint network environment is complex, existing large-scale different types of network data, causing difficult network intrusion virus classification. traditional virus feature extraction method is based on the single source network to determine the source of virus and make extraction of the virus feature, and in large joint network, due to the inability to determine the source of virus, leading to low convergence efficiency of virus feature extraction data classification model in joint network, and the feature extraction accuracy rate is very low [4,5], there are some disadvantages.

Theory of virus feature extraction in large joint network

Large joint network is to take the center network as the major and access to different types of protocol network, the diversification of network access type causes the multi-source data, for this kind of virus characters data extraction, it needs to identify the data source, anomaly data characteristics and variation feature extraction is the major, and the principle is as follows:

Assuming that the abnormal node features in the network are all the P dimensional feature vector, denoted $y=(y_1,y_2,\cdots,y_p)^T$, the matrix consisting of the P dimensional feature vector can be expressed as a $S_y=F[yy^T]$, the elements of which can be expressed as $\overline{y}=F[y]$. Abnormal behavior characteristic difference matrix of network can be expressed as $D_y=F\left[(y-\overline{y})(y-\overline{y})^T\right]$. The anomaly characteristics of network operation behavior is made extraction, also is to compare the network operation

behavior characteristics y with abnormal operation behavior characteristics $z=(z_1,z_2,\cdots,z_p)^T$, using the formula to describe:

$$z=V^Ty=(v_1,v_2,\cdots,v_p)^Ty=\begin{bmatrix}v_1^T\\v_2^T\\\vdots\\v_p^T\end{bmatrix}y \qquad (1)$$

Where, $z_j=v_j^Ty, j=1,2,\cdots,p$. It can determine y through formula (2):

$$y=(V^T)^{-1}=V_zz=V^Ty=(V^T)^{-1}z=V_z \qquad (2)$$

In the extraction of network virus features, first of all the weighted calculation is completed, so as to effectively extract the invading virus characteristics, weighted calculation formula can be used to describe as formula (3):

$$\hat{y}=\sum_{j=1}^{n}z_jv_j+\sum_{j=n+1}^{P}c_jv_j \qquad (3)$$

Through formula (4), it can find the error of network virus feature extraction:

$$\phi^2(n)=F[(y-\hat{y})^T(y-\hat{y})]=\sum_{j=n+1}^{P}F[(z_j-c_j)^2] \qquad (4)$$

if $c_j=F[(z_j)]=v_j^TF[y]=v_j^T\overline{y}$, the error of extracting network virus features reaches to a minimum, and:

$$\phi^2(n)=\sum_{j=n+1}^{P}F[(z_j-c_j)^2]=\sum_{j=n+1}^{P}v_j^TD_yv_j \qquad (5)$$

The process of extracting network virus features can use formula (6) to describe:

$$K=\sum_{j=n+1}^{P}[v_j^TD_yv_j-\mu_j(v_j^Tv_j-1)] \qquad (6)$$

Among them, μ_j is he network anomaly characteristics matrix, v_j is its' corresponding characteristics component.

Virus feature extraction model in large joint network

The basic model of virus invasion in large joint network

The ways of virus invasion in large combined network mainly includes the intranet access intrusion and external access intrusion. In large combined network, virus intrusion classification and extraction model use server to obtain the access data of the corresponding network, based on attribute optimization model of data access, to get the classification optimization model. Corresponding virus classification model can be described as:

$$W=H(x,x)-2\sum_j a_jH(x_j,x)+\sum_{i,j}a_ia_jH(x_j,x) \qquad (7)$$

$$f(x)=zqm(W-(H(x,x)-2\sum_j a_jH(x_j,x)+\sum_{i,j}a_ia_jH(x_j,x))) \qquad (8)$$

Among them, H is used to describe the identification attribute of the collected access data, W is used to describe the access data in the joint network, a is used to describe the attributes of access data. $f(x)$ is used to describe the classification parameter by using virus feature extraction model in joint network.

Network virus feature extraction algorithm based on unconstrained clustering correlation

If q is used to describe the number of abnormal nodes, uh is used to describe the $h(h=0,1,\cdots,q)$ -th interference attribute. Assuming that the sub interference characteristics in large joint network can be divided into G classes, $(v_1,v_2,\cdots v_n)$ is used to describe the corresponding sub interference attribute, the number of sub interference is n , each sub interference is expressed as $uh(vh_1,vh_2,\cdots,vh_n)$. Among them, $h=0,1,\cdots,n$, $uh(vh_1,vh_2,\cdots,vh_n)$ is used to describe all the interference nodes in joint network, unsupervised clustering correlation algorithm mainly through a variety of differences attributes

complete classification of several sub interference ^{uh}. Sub interference ^{d_j} belongs to the probability of degree of correlation in a certain category ^{g_k} which can be expressed as $^{z(g_h \bullet d_j)}$, and:

$$z(g_h \bullet d_j) = (z(g_h) + z(g_k \bullet \mu)) / z(d_j) \quad h = 0,1,\cdots,q \qquad (9)$$

Among them, $^{z(d_j)}$ is used to describe the prior probability of node anomaly detection in large joint network; $^{z(g_h \bullet d_j)}$ is used to describe the priori conditional probability of node classification. If the self-interference is the same property, the stability of the test results of sub interference probability $^{z(g_h)}$ is high, on the other hand, that is to establish the following link:

$$z(g_h \bullet d_j) = z(v_1 \bullet d) + z(v_2 \bullet d) + \cdots + (v_m \bullet d) = \sum_{h=2} z(v_h \bullet d) \qquad (10)$$

$$z(d_j) = \sum_{h=2}^{q} z(g_k) / z(d_j \bullet g_k) \qquad (11)$$

then getting the parameter properties, the following formula is used to determine the correlation probability which does not belong to the same interference characteristics of large joint network node:

$$z(d_j \bullet g_k) = R(q(d_j) + g_h) / R_{vh}^2 \qquad (12)$$

In formula (12), $^{R(q(d_j)+g_h)}$ is used to describe the number of interference contained in the joint network in the class ^{g_h}. $^{R_{vh}^2}$ is used to describe the number of nodes of interference attribute in the sample which will be detected. If zsin alcosa<b, then the corresponding node is the interfering node, otherwise it is non-interfering node. Through the above method, can realize the effective classification of multi feature interference of large joint network virus.

Using the repeated game factor optimize classification model

In this paper, in order to further make optimization of the analyzed multiple attribute characteristics of network virus classification process, in computing the probability of network virus characteristic data, the game factor is integrated into the constraints calculation process. The optimal function of data segmentation is reflected as shown in formula (13):

$$DZ_j = [\frac{D}{G_j + B_j} - \frac{P_j}{i_{jh}}] \qquad (13)$$

In the formula, $^{U_i = \lambda^2 + \sum q_i g_{ij} + q_0 g_i}$, $^{\max\{e, \min\{x, \beta\}\}}$ can be described by $^{[x]_e^\beta}$, D represents the constrained parameters. s and e represent the vector and threshold respectively.

Formula (14) is adopted to get the impact factor of game, in the process of the game, the step size is $^{\beta(s)}$, ^{s} is used to describe the maximum number of iterations:

$$\lambda(t+1) = \max\left\{0, \lambda(t) + \alpha(t)\frac{\partial U}{\partial \lambda}\right\} \qquad (14)$$

The step size of game is $^{\alpha(t) = \frac{\Delta p d}{P_0 g_{0m}\sqrt{t}}}$, ^{d} is used to describe the restriction factor of grid virus data classification, is generally the normal number. Formula (14) is transformed:

$$\eta(s+1) = \max\left\{0, \eta(s) + \frac{e}{\sqrt{s}}(\frac{\Delta q U_u}{q_0 f_{0m}} - 1)\right\} \qquad (15)$$

Using the optimal reaction function complete virus data classification in large-scale joint network, first of all, it needs to adopt the formula (16) to determine ^{E_i} and ^{S_i} of the optimal reaction function in joint network.

$$\begin{aligned} E_i &= \sum \frac{CZ_q k_q}{U_q(U_q + Q_q k_{qq})} \\ S_i &= \sum_{kfU} \gamma f_{iu} \end{aligned} \qquad (16)$$

Among them, $^{Z_q = q_q k_{qt}}$ is used for describing the parameters in large joint network virus data classification.

From the above analysis process, it can be summed that in the large joint network, the specific flow of grid virus data classification algorithm based on the repeated game theory is:

(1) The use of formula (17) complete the parameters initialization involved in joint network virus data classification:

$$q_s(1)=0, s=1, \gamma(1)=0, p \in P, s=1 \qquad (17)$$

(2) The use of formula (18) adjust the joint network data:

$$E_s(s+1)=\sum_{pfS}\frac{CZ_q(s)k_{sp}}{U_q(U_q+Z_q(s))} \qquad (18)$$

(3) The use of formula (19) regulate the impact factor of game:

$$\gamma(s+1)=\max\left\{0,\gamma(s)+\frac{e}{\sqrt{s}}\left(\frac{\Delta qU_q}{q_0 f_{0q}}-1\right)\right\} \qquad (19)$$

(4) The use of formula (20) adjust the parameters involved in the optimal reaction function:

$$S_s=\sum_{pfs}\gamma(s+1)f_{iq} \qquad (20)$$

(5) The use of formula (21) can get virus data classification parameters in large joint network:

$$Q_s(s+1)=\left[\frac{C}{E_s(s+1)+S_s(s+1)}-\frac{U_s}{k_{ss}}\right]_0^k \qquad (21)$$

(6) If the relevant data of data classification are accordance with formula (22), there are $s=s+1$, at the same time, reopen steps (2), whereas the end of iteration.

$$\frac{Q_s(s+1)-Q_s(s)}{q_s(s)}>\tau \text{ or } \frac{q_0 f_{0m}}{U_q}<\Delta q \qquad (22)$$

The above analysis process aims at virus feature extraction model in large joint network, according to the improved model of identification property of access data, using virus detection algorithm with unconstrained clustering correlation make accurate classification for network multi-feature interference. The constraint calculating game factor is integrated into classification probability calculation. Data game is used for filtering several probabilistic contrast in the process of network virus characteristic probability matching constraints, through the game constraint calculates the optimal reaction function, which makes the virus feature classification in large joint network to achieve optimal.

Simulation Study

Simulation data source

DARPA virus evaluation data set of Lincoln Laboratory in MIT are used as the sample data of simulation experiment. This paper adopts 15% data from DARPA data as the experimental data, and the data are random divided into training data sets and test data sets.

The evaluation index of the model

The two kinds of evaluation index of false alarm rate and detection rate are used in this experiment to assess the performance of the extraction model in this paper and traditional extraction model. In the experiment, the training set is made principal component analysis, get to get the result shown in figure 2

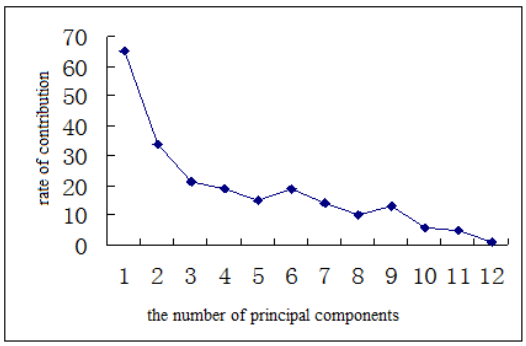

Figure 1 Pareto diagram of Principal Component

Experiment results and analysis

From the data in Table 1, it can be seen, the accuracy of average classification extraction using the proposed virus feature extraction model is higher than traditional extraction model. At the same time, error detection rate of the extraction model in paper is lower than the traditional extraction model, illustrating the use of this research model for virus features classification in large joint network, can eliminate invalid information in the data, improving the efficiency and the accuracy of network virus feature extraction.

Table 1 the statistical results of the performance by using two extraction models

attack types	false drop rate of traditional model	extraction efficiency of traditional model	false drop rate of the proposed model	extraction efficiency of the proposed model
DOS	10	92.9	5.4	95.7
R2L	12.9	91.7	7.7	93.9
PROBE	9.7	93.6	3.8	97.5
U2R	64	72.7	22.7	79

Conclusions

A kind of virus feature extraction model for large joint network based on unconstrained clustering correlation and repeated game factor is proposed in this paper, according to the identification attribute of access data perfect it. Using unconstrained clustering correlation virus detection algorithm make accurate classification of the multi feature interference in joint network. In the classification probability calculation, constraints computational game factors are introduced. Using data game filter multi-time probabilistic contrast in the features probability matching process of joint network virus. By calculating the optimal reaction function, makes the joint network virus feature extraction to achieve optimal. Simulation results show that, the proposed model can effectively extract the characteristics of joint network virus, and the efficiency and the accuracy is better than the traditional model, has obvious optimization effect.

References

[1] Xiao Haijun, Hong Fan, Zhang Zhaoli, Liao Junguo. Research on intrusion detection based on fusion classification and support vector machine [J]. Computer simulation, 2008, 25 (4): 130-132.

[2] Zhang Fengbin, Yang Yongtian, Jiang Ziyang. Application of genetic algorithm in intrusion detection based on network anomaly [J]. Acta Electronica Sinica, 2004, 32 (5): 875-877.

[3] Tian Junfeng, Zhang Jing, Bi Zhiming. The study of intrusion detection based on improved RBF neural network [J]. Computer engineering and application, 2008, 44 (31): 135-138.

[4] Zhang Xueqin, Gu Chunhua. A network intrusion detection feature extraction method [J]. Journal of South China University of Technology: Natural Science Edition, 2010, 38 (1):81-86.

[5] Sun Dapeng. The research of improved fuzzy clustering algorithm applied in intrusion detection [J]. Computer and digital engineering, 2010, 38 (3): 88-91.