A Prediction Model of DoS Attack's Distribution Discrete Probability

Wentao Zhao, Jianping Yin, Jun Long
School of Computer Science
National University of Defense Technology
Changsha 410073, Hunan Province, China
zhao_bruce@sina.com

Abstract

This paper describes the clustering problem first, and then utilizes the genetic algorithm to implement the optimization of clustering methods. Based on the optimized clustering on the sample data, we get various categories of the relation between traffics and attack amounts, and then builds up several prediction sub-models about DoS attack. Furthermore, according to the Bayesian method, we deduce discrete probability calculation about each sub-model and then get the distribution discrete probability prediction model for DoS attack.

1. Introduction

The process of prediction analysis is a process of using some method or technology to explore or stimulate some unknown, undiscovered or complicated intermediate processes based on previous and present states and then speculated the results [5].

In an early warning system, accurate prediction of DoS attacks is the prime aim in the network offence and defense task. Detection based on abnormity is effective to detect DoS attacks. A various studies focused on DoS attacks from different respects [2][6][10]. However, these methods required a priori knowledge being a necessity and were difficult to discriminate between normal burst traffics and flux of DoS attacks. Moreover, they also required a large number of history records and can not make the prediction for such attacks efficiently.

Based on data from flux inspecting and intrusion detection, we propose a prediction model of DOS attack's distribution discrete probability based on clustering method of genetic algorithm and Bayesian method.

Due to various interference factors, the frequency of the DoS attack is considered to be a random variable. And probability is an effective way to describe randomness. Proper attack prediction models should not only identify the amount of possible future attacks but also provide the distribution information, which is the probability distribution of attacks in a certain interval.

Using the pattern classification method from [3], the paper views input and output relation of DoS attack as a multi-pattern random variable. Adopting the genetic algorithm to implement the optimization of clustering methods, the paper presents a classification method for DoS attack data, which means getting the various categories of the mapping relation between the traffic and attack amount and building up several prediction sub-models for Dos attack. Meanwhile, the paper deduces a discrete probability calculation method about each sub-model and gets the probability distribution of DoS attacks in a certain interval for any future time point.

The rest of the paper is organized as follows. Section 2 describes the clustering analysis in our prediction model. Section 3 provides the genetic optimization algorithm in the prediction model. Section 4 introduces the input and output of the prediction model. Section 5 presented the clustering of Dos attack's sample data and the building of sub-model. Section 6 gives the Bayesian method for the prediction model. Section 7 makes the conclusion.

2. Clustering Analysis in Prediction Model

Clustering analysis is a unsupervised classification method, which is used to obtain data distribution or to preprocess data without a pre-assigned class [7]. A good clustering method results in high quality clusters. Whether or not the clustering result is good depends on the similarity evaluation approach for the clustering method and the way to implement that approach [9].

The aim of clustering methods is to search in sample space a group of best lines, to which the distances is minimum from sample points . The input and output mapping relation of DoS attack is nonlinear to some extent. Unfortunately, the classical nonlinear model of input and output is complicated and difficult to express in parameter forms.

978-0-7695-3185-4/08 $25.00 © 2008 IEEE
DOI 10.1109/WAIM.2008.87

The paper adopts a special clustering method, which takes linearization into consideration, to properly divide the input and output mapping relation of DoS attack, that is to divide the input space into some subspaces and to use multiple linear regression to fit the input and output relation of the sub samples. Since the sub samples have a better linearization, the fitted linear model is of high accuracy.

Definition 1. In the input space X^{m+1}, the vertical distnce from point $x_i = (x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$ to line $y = a_1x_1 + a_2x_2 + \cdots + a_mx_m + b$ is

$$dist = \frac{|a_1x_{i1} + \cdots + a_mx_{im} + b - y_i|\sqrt{x_{i1}^2 + \cdots + x_{im}^2}}{\sqrt{(a_1x_{i1} + \cdots + a_mx_{im})^2 + x_{i1}^2 + \cdots + x_{im}^2}} \tag{1}$$

Choose some lines, and classify data points into different categories according to their distances to the chosen lines.

Suppose that a group of c lines $l_j, (j = 1, 2, \cdots, c)$ are chosen, then calculate the distance $dist_{ij}$ from data points to the group of lines respectively. If the distance from ith point to kth lines satisfies $dist_{ik} = min_{j=1,\cdots,c}dist_{ij}$, then classify ith point to kth category. In the same way, all the data points in the sample space are divided into categories they belong to.

Definition 2. Suppose that after classification, there are C_j data points, $x_i = (x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$, in the jth category, $i = 1, 2, \cdots, C_j$, and the corresponding line is $y = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jm}x_m + b_j$, then distance from each data point in the category to the line is $dist_{ij}$,

$$dist_{ij} = \frac{|a_{j1}x_{i1} + \cdots + a_{jm}x_{im} + b_j - y_{im}|\sqrt{x_{i1}^2 + \cdots + x_{im}^2}}{\sqrt{(a_{j1}x_{i1} + \cdots + a_{jm}x_{im})^2 + x_{i1}^2 + \cdots + x_{im}^2}} \tag{2}$$

$i = 1, 2, \cdots, C_j$. And distance from each data point in the category to another line $k = 1, 2, \cdots, c, k \neq j$ is

$$dist_{ik} = \frac{|a_{k1}x_{i1} + \cdots + a_{km}x_{im} + b_k - y_{im}|\sqrt{x_{i1}^2 + \cdots + x_{im}^2}}{\sqrt{(a_{k1}x_{i1} + \cdots + a_{km}x_{im})^2 + x_{i1}^2 + \cdots + x_{im}^2}} \tag{3}$$

$i = 1, 2, \cdots, C_j$. And distance from jth line to the other lines can be obtained by calculating the mean distance from the center of jth category to each lines.

$$\overline{dist}_{jk} = (\sum_{i=1}^{c_j} dist_{ik})/C_j \tag{4}$$

$k = 1, 2, \cdots, k \neq j$.

The aim of clustering is to obtain the minimum distance from sample data in the category to the line the category belongs to while maintain distances between lines as far as possible. Therefore, we can define the object function for classification.

Definition 3. The objective function of classified clustering is the ratio of mean distance from data points in the category to other lines to that from the data points in the category to the line the data points belong to.

$$F = \frac{\sum_{j=1}^{c}[(\sum_{k=1,k\neq j}^{c} \overline{dist}_{jk})/(c-1)]}{\sum_{j=1}^{c}(\sum_{i=1}^{C_j} dist_{ij})/C_j} \tag{5}$$

According to the description above we know that the data points in sub categories are near the corresponding line and the linearization of data points within a category is much better compared to before, which makes multiple linear regressions of data points in sub categories more accurate. Due to large space dimension of lines and various parameters of the objective function in researching for optimization, routine optimization methods such as the gradient descent method are incapable to deal with the problem, so the genetic algorithm is used to optimize the cluster problem.

3. The Genetic optimization algorithm in the prediction model

The genetic algorithm is an algorithm that imitates the response mechanism of the life-form immune system and the iterative optimization mechanism of populations [1]. This algorithm starts from an initial antibody population, through operators such as inhibitory concentration control, memory, crossover and mutation, and then it would come up with an overall optimal solution with the ability to cope with parallel search [8]. The concentration adjustment method of the Immune Regulation can keep the diversity of antibody populations well [4] to raise the ability of overall and partial search in this algorithm. Meanwhile, the immune memory function can insure that it will come up with an overall optimal solution fast.

The classification quality not only depends on the selection of the classifying lines, but also relates to the amount of the classifying lines. The length of chromosome doesn't change in the evolving process of the traditional genetic algorithm. In clustering problems, it means that the amount of category can not change. In order to improve this problem, consulting the dynamic immunity evolvement algorithm in reference [11], we take the crossover operator based on truncation and stitching in order to change the amount of chromosome. Then it can optimize both the amount and the method of classification.

Considering the clustering problem with data points in multi-dimension space, antigens corresponds to the quality of clustering methods with data points, viz the objective function for classification in definition 3; while antibodies corresponds to the various clustering methods of data points, viz the lines chosen for classification.

This paper codes the antibodies by floating-point coding method. Suppose that there are c_i categories of data points in the ith clustering method (there are different categories

in different classification method), there are parameters of c_i lines composing the ith antibody, viz.

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2m} & b_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{c_i1} & a_{c_i2} & \cdots & a_{c_im} & b_{c_i} \end{bmatrix}, (i = 1, 2, \cdots, s)$$

s is the size of the populations. The fitness function is defined as the objective function of the clustering problem in definition 3.

When using the genetic algorithm to cluster data points, the first step is generating an initial antibody population randomly, viz choosing c_i lines randomly for the ith antibody and classifying the data points into c_i categories by the distance between these points and these lines. Second, update the antibody populations, namely refresh the parameters of the c_i lines, including the refresh of memory antibody base, antibody memorizing, antibody choosing, antibody crossover and antibody mutation based on the concentration Inhibition. The refresh of the antibody base takes the purpose that finding out the antibodies that have the highest fitness in every generation of the antibody populations. If the degree of fitness is higher than the old memory antibody then replace it, otherwise keep the antibody base unchanged. The antibody memorizing is used to find out the antibodies that have the lowest fitness in the antibody populations and then replace them with the best memory antibodies. For more information about the operation of antibody choosing, antibody crossover and antibody mutation, please refer to [10][11].

4. The input and output of the prediction model

In the prediction model for DoS attacks, we take network traffic in the recent history and anticipated time as a part of the input of the model. Considering the dynamic response characteristic of the prediction model, we should also take the amount of the actual DoS attack in recent history as the input of the model in order to keep the balance of the attack level in short time.

The whole input of the prediction model for DoS attack includes network traffic in anticipated time $Traffic_{Anticipated}$, network traffic in recent history $Traffic_{History}$ and actual DoS attack amount in recent history $DoS_{History}$. The output is the DoS attack amount in anticipated time $DoS_{Anticipated}$. Suppose that the input is the network traffic and the actual DoS attack amount in recent m periods of time, and the amount of sample points is n. Then the input of the model is a matrix with $n \times (2m + 1)$ dimension, and the output is a vector with $n \times 1$ dimension.

5. Clustering of Dos attack's sample data and the building of sub-model

With the influence of many factors, the input-output relation between network traffic and attack amount is very complicated. When analyzing and training a large number of history sample data, it is difficult to express the input-output relation exactly with single model and the precision is not high enough.

With the clustering method of the genetic optimization algorithm, the sample data are first classified into different categories according to the linear relation they satisfy with. Then linear sub-model is built through the multiple linear regressions of the classified sub-class sample data to implement the complete fitting of the complicated input-output relation of the sample data.

First, cluster the large number of the history DoS attack's input-output sample data according to the clustering method based on genetic optimization algorithm to get the appropriate partition of the prediction sample data of DoS attack amount in linear space and the corresponding DoS attack's input-output relation in different possible models. Then, coordinate the clustered sub-class sample data and adopt the multiple linear regressions based on the Least Squares Principle to get the linear input-output relation of each sub-class sample data which is the sub models of the prediction model of DoS attack's distribution discrete probability. As the clustered sub-class sample data are well linearized, the multiple linear regressions can get a high precision.

6. The prediction of Dos attack's distribution discrete probability based on Bayesian method

Bayesian decision-making method is a basic method in statistic pattern recognition. Suppose the sample space of experiment E is S, A is an event of E and $B_1, B_2, \cdots, B_n$ is a partition of S. As $P(A) > 0$ and $P(B_i > 0), (i = 1, 2, \cdots, n)$, then

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}, i = 1, 2, \cdots, n \quad (6)$$

The essential of the above equation is converting the prior probability of each category to the posteriori probability with Bayesian formula according to conditional probability. As in this paper, it is to convert the prior probability of each sub-class sample to the corresponding posteriori probability via observing the distances between the prediction point and each sample point.

The prior probability of each class is acquired by calculating the ratio of the count of each class of training sample

points C_j to the total sample count T:

$$P(B_j) = \frac{C_j}{K}, (j = 1, 2, \cdots, c) \tag{7}$$

Definition 4.: Suppose the predicted input sample is $(x_1, x_2, \cdots, x_m)$ and this input-output relation belongs to some sub-class conditional probability $P(x|B_j)$, which can be defined by the vertical distance $dist'_j$ between the predicted point of the jth sub-model and the corresponding line of the model:

$$dist'_j = \frac{|y'_j - (a_{j1}x_1 + \cdots + a_{jm}x_m - b)|\sqrt{x_1^2 + \cdots + x_m^2}}{\sqrt{(a_{j1}x_1 + \cdots + a_{jm}x_m)^2 + x_1^2 + \cdots + x_m^2}} \tag{8}$$

$$P(x|B_j) = \frac{1/(dist'_j)^\beta}{\sum_{j=1}^{c}[1/(dist'_j)^\beta]}, j = 1, 2, \cdots, c \tag{9}$$

Where $y'_j = a'_{j1}x_1 + a'_{j2}x_2 + \cdots + a'_{jm}x_m + b'_j$, $a'_{j1}, a'_{j2}, \cdots, a'_{jm}, b'_j$ is the parameter of the jth sub-model and $a_{j1}, a_{j2}, \cdots, a_{jm}, b_j$ is the parameter of the line of the jth sample in linear clusters. Because the values of actual distances are rather small and their subtle difference may induce large difference of conditional probability, exponent β is used to correct the model.

In section 3, several linear prediction sub-models are achieved with the optimal clustering mode. With these sub-models, we can obtain the possible corresponding amount of the output DoS attack for an input predicted point. Then according to the Bayesian probability prediction, we can obtain the probability of all these possible amounts of the DoS attack. The discrete probability distribution composed of these series of attack amount and probability is the final output prediction result of the prediction model in this paper.

7. Conclusion

This paper begins with the relation exists between network traffic data and the amount of DoS attack, and then proposes a clustering method based on the genetic optimization algorithm to implement the classification of DoS attack data. This method first gets the proper partition of the relation between the network traffic and the amount of DoS attack based on the optimized clustering and builds the prediction sub-models of DoS attack. Meanwhile, with the Bayesian method, the calculation of the output probability corresponding to each sub-model is deduced and then the distribution of the amount of DoS attack in some range in future is obtained.

References

[1] F. D. B. *Evolution computation: toward a new philosophy of machine intelligence*. Piscataway: IEEE Press, 1995.

[2] J. B. D. Cabrera, L. Lewis, X. Qin, W. Lee, R. K. Prasanth, B. Ravichandran, and R. K. Mehra. Proactive detection of distributed denial of service attacks using mib traffic variables - a feasibility study. In *Proceedings of International Symposium on Integrated Network Management*, 2001.

[3] P. B. L. Ernan Ni. Forecasting power market clearing price and its discrete pdf using a bayesian-based classification method. *IEEE Trans. on Power Systems*, pages 1518–1522, 2001.

[4] G. Jian. Cluster analysis based on c-means and immune genetic algorithm. *Computer Engineering*, 29(12):65–66, 2003.

[5] G. R.-K. T. K.-R. Muller, S. Mika and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Transaction on Neural Networks*, 12(2):181–201, May 2001.

[6] A. B. Kulkarni and S. F. Bush. Detecting distributed denial-of-service attacks using kolmogorov complexity metrics. *J. Network Syst. Manage.*, 14(1):69–80, 2006.

[7] L. C.-J. Lin Kuan-Ming. A study on reduced support vector machines. *IEEE Transactions on Neural Networks*, 14(6):1449–1459, 2003.

[8] B. S. Maulik U. Genetic algorithm-based clustering technique. *Pattern Recognition*, 33(9):1455–1465, 2000.

[9] J. Raychaudhuri.S., Stuart and R. Altman. Principal components analysis to summarize microarray experiments: Application to sporulation time series. In *In Proc. Pacific Symposium on Biocomputing*.

[10] S. Z. Z. X. Sun Qindong, Zhang Deyun. Detection of distributed denial of service attacks based on flow connection density. *Journal of Xi'an Jiaotong University*, 38(10):1048–1052, 2004.

[11] L. Y.-g. ZHU Hong-xia, SHEN Jiong. A novel dynamic clustering algorithm and its application in fuzzy modeling for thermal processes. *Proceedings of the Chinese Society for Electrical Engineering*, 25(7):34–40, 2005.