A Prediction Model for Network Traffic Anomaly Detection

ZHANG Yong-ping, QI Zhi-wei, LIU Jia

Dept. of Computer Science and Technology, China University of Mining & Technology, Xuzhou Jiangsu 221116, China

Keywords: Prediction Model, Traffic Detection, Network Security, Intrusion Detection, Anomaly Traffic

Abstract

In order to improve the defects in intrusion detection system (IDS) on network traffic anomaly detection (NTAD). A new IP based NTAD was proposed, which connected with linear regression analysis and IP adjust. This mechanism point at the features of the network, use linear regression analysis on the IP data, this can predict the abnormal traffic more accurately. The experiment results also showed the new mechanism have more accuracy and higher recall ratio.

1 Introduction

With the widely spread of Internet, the network security has become more and more important: virus, worms and DoS attacks threat us everyday [1]. An ideal method is to build a completely safe system. But in that case, the system required every user must be identify and recognized, encryption technologies and access control policies must be used to protect the data. That is impossible by now. A practical approach is to build a safety system which is easy to implement, at the same time establish a considerable support system such as IDS [6].

IDS can find out whether there are violations of security policy behaviours or signs of attacks in the network, through the analysis on the information. Its key is to describe the behaviours of network traffic, to find the possible abnormal signals or determine the new network threats, then warn the administrator or take the initiative for responding [5]. At the beginning, the Full Collection is applied to detect the threats by IDS, it is impossible for the rapidly expanded size of network. So we apply data sampling as the data resource for the anomaly detection [2], but sampling is not 100% accurate to the overall flow, it may affect the results [3].

To solve the problem traffic prediction model can be a good solution. Prediction model analysis the existing transactions by time order to create some rules. When the traffic is not in conformity with the rules it is invasion. Poisson and Markov model [7] in traditional queue theory were used at early time, however, during the network development, it is insufficient to predict the strong sudden and non-stability network traffic with a linear model [4], so the subject that how to transform the non-linear problem to a linear problem which can be analysed easier, is a very hot research area. The weighted sampling algorithm proposed here is an attempt to solve the question.

2 Network Traffic Prediction

Network anomaly means the data deviate from their normal behaviors. There can be many reasons for it: sudden large density access, hacker invasions, network device faults or wrong operations. Those factors happen suddenly and attack the network and computers in a short time. However the anomalies can be found in advance by monitoring the network traffic, so how to detect the anomaly in traffic and corresponding rationally is very important factor to keep the network perform efficiently.

The predict process can be described by follows [1]:

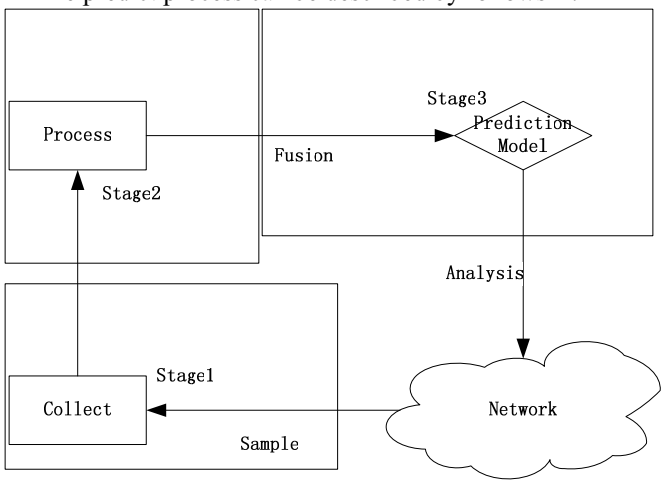


Figure1: Prediction model for network traffic

Stage1 collect: including the configuration of network, dynamic data of transactions. Mainly method is sampling, there are different ways to sample, first is sniffer, but without specific equipment, the data can be inaccuracy. Another way is to get the SNMP information from the routers, the data can be impartial but it also stressed the network system. Reasonable sampling techniques can reduce the operations on traffic data; this can lower the cost of routers' CPU, RAM and bandwidth. At present, the most widely used method is the random sampling based on IP data.

Stage2 process: deal with the data collected from last stage, to find out useful information, especially the abnormal information, such as higher delay, larger CRC error and so on. There are lots of methods: mathematical statistics, machine learning and data mining analysis, etc.

Stage3 fusion: analysis information gathered last stage, judging the reason for the troubles, construct a traffic model in universal, the methods are mainly on artificial intelligence.

With the 3 steps, a non-linear model can be approximately simplified to a linear model which is easier to be processed.

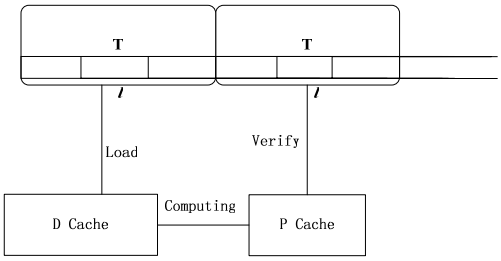
3 Traffic prediction base on IP

3.1 Prediction model

The IP weighted prediction model is proposed according to the own characteristics of the network. System detects and analyses the IP data packets, then build the prediction model through curve fitting and parameter estimation.

Shown in Figure2, system compute with the last m packets, obtain n group linear regression parameters *(a, b)*, keep the parameters in P cache, reserve the last packet in the D cache, and then process with following data. The size of sliding window T is set to *l*, it means *l* packets will be analysed in one time, and when the cache is full, and it will be cleared, and read the next group data. As it goes on the data can be analysed continuously.

For the next, Substitute the parameters in P cache into regression function compared the result with data in D cache, if they are not equal, fix the parameters until the result reach in the error range.



Fiugrue2: Prediction Model Build

3.2 Prediction on IP weighted

Conception: To be simplified, the model only consider with the time factor t. If the width of sliding window is l, according to the observed value Y(y_i,y_{i+1},.....y_{i+l-1}) in current sliding window T(t_i,t_{i+1},.....t_{i+l-1}), a regression function y=*f(t)* about current data is able to be calculated. Then the function can be used to predict the IP data after time t_{i+l-1}, compared with the actual packets, we can get the predict value fits the error range.

The prediction model is obtained through the data packets in current cache, so it may not reflect the whole network's situation. To improve this, IP address can be set to a special weight value, it is used to adjust regression parameters we got before, to clear the potential error.

Algorithm: Prediction model
Start
1 Set T=Null; //initialization
2 for i=1;i++;i<n;
{
for j=i;j++;j<i+l;
3 get $y_j \in T$; //read data
4 a_i , b_i ←*Linear Regression (T) ;*
//linear regress the data, get parameters a,b
5 cache=get (y_i) ;
//keep the data in DCache
6 Par=get(a_i,b_i); //keep the data in P Cache
for k=1;k++;k<m;
7 {get_cache_ip(y_{k-1},y_{k-2}.....,y_{k-m+1})
//get the packet in cache
For i=1;i++;i<m;
8 $w_{k-i}=ip_{k-i}$ ***xor*** ip_k;
//set weight value
9 $a_k=w_{k-1}a_{k-1}+w_{k-2}a_{k-2}.....w_{k-m+1}a_{k-m+1}$;
//adjust the parameters
10 set uk=f(ak,bk,tk); //Predition
11 Get yk $\in T$; //get actual data
12 Rss=get_rss(uk,yk); //compare
}
}

According to the algorithm, with a linear regression process the non-linear real data can be simplified to a linear model. The data packet with closer IP addresses to the original node will be set to higher weight, so it can effect the prediction greater. Also the parameter can be fixed by circumstance. The IP weighted algorithm uses multi-parameters to fix the prediction, so it can eliminate the interference caused by data anomaly. The key point is connecting the IP address with the weight.

4 Performance experiment

4.1 Environment

The experiment use a tcp connection recorded in 30 days as a data resource, and use mathematic tools to complete the IP weight. A computer with Windows XP, 1.69GHZ CPU and 1G RAM is used as the equipment.

Draw 2000 data packets from data resource to prediction, set *l* to 250, θ to 150, and they are fixed continuously, until obtain the result in error range, then apply the model to predict the data packets. As a contrast, a traditional linear regression also is used to predict.

Packet Size	Delay	IP
2.427995	1.24383	128.97.154.3
2.803033	3.96513	128.8.142.5
4.817224	1.02839	140.98.2.1
12.254572	138.168	128.49.4.103
17.478016	10.0858	128.32.133.1
33.453963	2.16477	192.48.232.17
.....	.....	.....
36.735788	13.1779	128.233.1.12

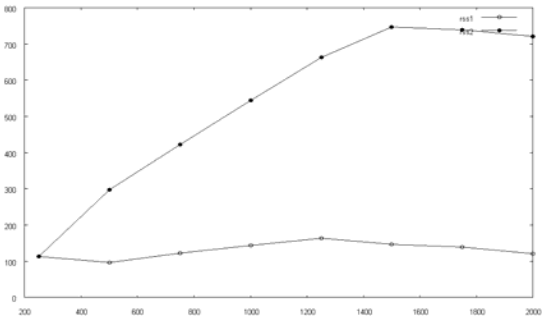
Table1: Data resource

4.2 Experiment

Prediction model: Figure3 shows the accuracies for two prediction models, white "O" marks the IP weighted prediction error while black "O" as the traditional. We can see the residual error (rss) that created by tradition prediction model is much higher. It is clear that the IP weighted prediction is more accurate.

Load Rate	Total	IP weighted prediction				Traditional prediction			
		Find	Loss	Error	Acc	Find	Loss	Error	Acc
18%	37	37	0	0	100%	37	0	0	100%
52%	37	38	0	1	99.7%	30	7	0	81.0%
85%	37	40	1	4	94.6%	18	20	1	45.9%

Table2: Detection Result



Fugure3: Accuracy Comparison

System resource usage: Figure4 shows the resource usages of the two prediction model building, keep the same, white "O" marks the IP weighted prediction and black "O" for traditional. The IP weight algorithm has to load and process more packets so the usage is slightly higher, but they were generally equal. Compares to the accuracy the high system resource cost by IP weighted prediction model is acceptable.

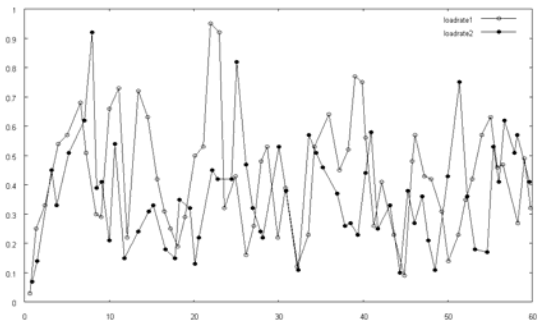


Figure4: Usage of resource

Detection Result: We create some abnormal traffic, to verify the performance by compare the detection rates. In the experiment, the simulation environments are 3 situations with network load rate at 18%, 52% and 85%. Table2 is the record list. We can see at the lower load rate, there are no performance contrast between the traditional the IP weighted prediction. But with the increasing load rate, IP weighted prediction can keep the performance in high level while the traditional model decreased. When the network is very busy, the load rate is higher than 85%, the traditional prediction model can only find less than half of the errors while the IP weighted prediction model still can find more than 90%.

5 Conclusion

A new prediction model apply in IDS is proposed in this paper. Based on the experiment result, the new model can predict the network traffic more accurately. Although it is more complex and the resource usage is higher, compare to the accuracy increasing, the cost is acceptable. The result shown, IP weighted prediction model can detect the anomalies more accurately and completely, and it has more stable performance

Acknowledgements

The acknowledgement for funding organizations etc. should be placed in a separate section at the end of the text.
Thank you for your cooperation in complying with these instructions.

References

[1] Bai-xian Zou. Detection and Prediction of Network Traffic Anomaly[D].Beijing: Graduate University of Chinese Academy of Sciences,2003.

[2] Choi B Y, PARK J, ZHANG Z. Adaptive Random Sampling for Traffic Load Measurement[C]. *Proceedings of IEEE International Conference on Communications. Piscataway, USA: IEEE,* **2003**: pp,1552-1556.

[3] Jianning M, Chen-Nee C, Ashwin S, Et a1.Is Sampled Data Sufficient for Anomaly Detection[C]. *Proceedings of the 6th ACMSIG COMM on Intemet Measurement. New York: ACM Press,* **2006:** pp,165-176.

[4] Lee M.Video traffic prediction based on source information and preventive channel rate decision for RCBR[J].*IEEE Transactions on Broadcasting,* **2006, 52(2):** pp,1-11.

[5] Qiao Pan, Chang-xing Pei, Chang-hua Zhu. Novel Traffic Sampling Method for Anomaly Detection[J]. *Journal of Xi an Jiaotong University,* **2008,2(42):** pp, 175-178.

[6] Ru-hui Ma.The Technology Research of network Security based on Network Traffic Anomaly Detection[D].Wuxi: Jiangnan University,2008.

[7] Wen-xian Jiang, Anomaly Detection Technology for Network Traffic Based on Random Fractal and Markov Model[J].*Communications Technology,* **2008.10(41):** pp,166-169.

[8] Zhi-jun Shen. Research on Security Information Fusion Model in Network Security Manager System[D]. Beijing: Tsinghua, 2007.