Analysis about Performance of Multiclass SVM Applying in IDS

Gang Zhao, Jianhao Song, Junyi Song

School of Information Management, Beijing Information Science & Technology University, Beijing, China

Abstract

This paper presents a novel network intrusion detection approach with the Support Vector Machine embedded in and K-fold cross-validation method compounded for optimizing the attributes and SVM model. Compared with some representative machine learning method, online data experimental results show that this method can be used to reduce the rate of False-Negatives in the intrusion detection system.

Keywords: IDS; SVM; machine learning; rate of False-Negatives

1. Introduction

Noting that network becomes closer with people’s working and living, the security issues of network also receive more attentions. Internet services have become essential to business commerce as well as to individuals. With the increasing reliance on network services, the availability, confidentiality, and integrity of critical information have become increasingly compromised by remote intrusions. An Intrusion Detection System (IDS) [1] monitors the events occurring at a host or within a network and raises alerts for suspicious ones. IDS are classified as Anomaly Based [2] or Signature Based [3] depending on the technique used for detecting intrusions. A statistical anomaly-based IDS determines normal network activity like what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other and alert the administrator or user when traffic is detected as anomaly.

Because of the inherent flaws existed in two methods mentioned above, rates of False Positive and False Negative are high. Thus, researchers introduce machine learning methods to solve the problems of data processing. However, most traditional machine learning algorithms are based on assumptions that amount of samples tends to infinity and required high data regularity. Until recently, various intelligent IDS cannot create ideal results [4].

This paper studies the small sample learning in the field of machine learning and adopts the Support Vector Machine (SVM) [5] to resolve the heavy overhead, slow detection rate and high rate of False-positives and False-negative problems in NIDS. Further, this paper presents methods to select and optimize the attributes and model of SVM. Moreover, this paper analyses the online data experimental results compared with some representative machine learning using port scan and DoS attacks software to emphasize validity of the method.

The rest of the paper is organized as follows. In Section 2, we present the typical intrusion detection system and related works. In Section 3, we present our method with SVM in detail. In Section 4, we show the results of experiments and analyze them. Finally, we present concluding remarks.

2. SNORT Intrusion Detection System and Related Work

Snort is a free and open source network intrusion detection system (NIDS) created by Martin Roesch in 1998, which is a powerful lightweight NIDS that has abilities to analysis real-time data, match content in network data and log. It can detect a variety of attacks and provide real-time alarm for those attacks, and runs on a host to monitor the network data. SNORT can match patterns between the network data and detection rules, thereby detecting a variety of possible intrusion attempts. Moreover, SNORT also has good scalability and portability.

Similar to some other traditional IDS, the development of SNORT has encountered a bottleneck. IDS face the following major problems. IDS have several commonly detection methods include feature detection, anomaly detection, state detection, protocol analysis, etc. These detection methods all have flaws. Such as anomaly detection commonly used statistical methods to detect, but it is difficult to effectively determine the threshold of statistical methods, small value will produce a lot of false-positives, big value will produce a large number of false-negatives. In protocol analysis detection methods, the normal IDS just simply dealt with commonly protocol such as HTTP, FTP, SMTP, etc. The large number of rest protocol packets entirely possible cause false-negatives. If consider supporting as many as possible of the protocol type analysis, the cost of the network will not be able to afford.

IDS can only identify the IP address, cannot locate the IP address and identify the data source. When IDS found the attack events, it can only close network exports and a few ports of server, but this close will also affect other normal users' use. Thus, it lacks a more effective response handling mechanism.

Existed IDS products are mostly used feature detection technology, these IDS products cannot adapted exchange technology and development of high-bandwidth environment, in the case of large flows impact and multi-IP fragmentation, IDS will paralysis or loss packages and then form DoS attacks.

3. The Proposed SVM approach with K-fold Cross-Validation in the IDS

In machine learning, Support Vector Machines (SVM) is supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. SVM is the machine learning technique developed in mid-1990s and popular statistical learning method in common use, which is based on structure risk minimization principle. When comparing with traditional neural networks method which based on empirical risk minimization, SVM not only has more simple structure, but also has better generalization ability for small samples. SVM method is based on VC Dimension theory and structure risk minimization principle of statistical learning theory, shown in Figure 1. With limited sample information, SVM can reach the best compromise between complexity and learning ability of model, in order to obtain the best generalization ability [6].

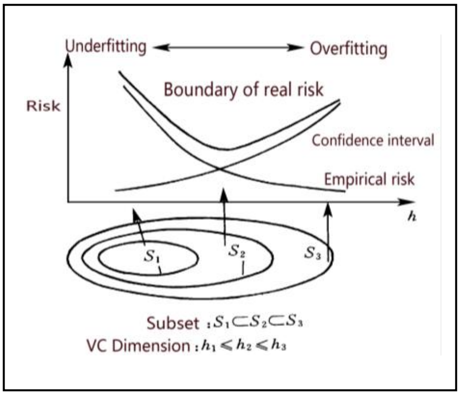

Fig.1 Structure risk minimization.

SVM is evolved from optimum classification surface in the linear separable case. Under linear separable case, it can converse issue of configure optimum hyper-plane to calculate the minimum value of $\Phi(w)= \| w \|^2$. Support vector interval can be calculated as $2/\| w \|$. Distance between any point x and Hyper-plane is $(w \times x+b/\| w \|)$, while the optimum hyper-plane which having a maximum interval need in such condition: the number of VC dimension of regulation hyper-plane should not bigger than min([R2,A2],n)+1, n is amount of dimensions of vector space, all vectors waiting for split are located within a hypersphere have radius R, and $\| w \| \leq A$. Build the optimum hyper-plane to split two types can be transformed to quadratic programming, that is calculate the minimum value of $w^2/2$ [7] when $yi(w \times xi+b) \geq 1 (i=1\ldots l)$.

The traditional solutions of false-positives and false-negatives in intrusion detection rely on improvements of analysis technology. Mainly analysis methods of intrusion detection include statistical analysis, protocol analysis, behavior analysis, etc. However, this method can only be applied in some types of attacks, such as DoS attacks, effect is not obvious for some attacks with small data scale.

Based on reorganization of network data stream, protocol analysis technique can understand application protocols then use pattern matching and statistical analysis techniques to ascertain attacks. If use protocol analysis, alarm will emerge only when event is detected coincide with the protocol like HTTP. Assuming this feature appeared in Mail, because they do not coincide with the protocol, it will not alarm.

Behavior analysis technology is not only a simple analysis for single attacks but also according to the latter and previous event to confirm whether the attacks occurred, whether the attack behavior is in force. However, because of extremely difficult of algorithm processing and rule-making, it is not mature enough yet.

With requirements of large amount of data processing, performance requirements for IDS are also gradually increasing, hence Gigabit IDS and other products emerged. However, if intrusion detection products not only have attack analysis, but also had the function of contents recovery and network audit, system can hardly working under a Gigabit environment completely.

IDS discover attacks then sent to firewall automatically, firewall load dynamic rules to intercept intrusion. This function called firewall interacting. It is not yet coming into practicality, primary a concept. Casual use can cause a lot of problems. It will make a negative impact on firewall stability and network applications if tested inadequately.

Above methods are not good solutions to the problems which IDS faced. So people put the focus on combination of AI methods and IDS. Most existing AI methods applied IDS are in offline testing stage, due to lack of a theoretical basis and methods itself have inherently flawed, those IDS cannot achieved ideal results. This study through compare several machine learning algorithms like Bayesian, neural networks, decision trees and SVM, found that SVM algorithm is an ideal methods for invasion judgment[8] because sample it requires is small, classification is accuracy and other features. Therefore, this study decided to use SVM algorithm as detection core.

There are two advantages by using SVM as solution: as a classification algorithm, SVM calculated the optimum solution of distance, which is the fairest classification. It can resolve high rate of false-positives and false-negatives problems exist in IDS. And, volume of model that SVM used for classification is small. Because using small samples to classify is an advantage of SVM, so it greatly reducing detection time compared to traditional IDS. Through a lot of tests, we verified that aiming at IDS, SVM model are better than neural network, Naive Bayes and other traditional classification algorithms [9][10] both on training and testing.

Because IDS will ultimately be applied in actual network environment, this experiment decided to use famous NIDS, SNORT, as framework and SVM algorithm as detection core, applied SVM in practical online testing. Until now, most SVM applications are testing in offline environment; it cannot be run in online environments, which are the problem our studies can solve.

K-fold cross-validation method can not only optimize model but also test result of attribute selection. Stronger characteristics attribute have, smaller model after optimize. Nevertheless, if characteristics of attribute are so strong that one attribute have decisive impact to entire result, it is not suitable for application of SVM algorithm. Because advantage of SVM algorithm is high-dimensional pattern recognition, if dependent on only a few attributes mean dimension is low, expert system will be more suitable in this case. So amount of attributes is not the only factor, information contained in characteristics of attribute should also similar. Only in this way could SVM algorithm most suitable.

According to a variety of different types of intrusion actions online, this study proposed that based on BSVM, through modify SNORT, we can implement multiple classifications of intrusion actions.

System decodes captured network flow from network card by Decode module, then transmits decoded data to Preprocessor module for preprocess, reassemble slice packets and unify format of URL string requested by HTTP. After preprocessing, data is converted into SVM acceptable format. If we want to train SVM model, system will pass data to SVM-train module, model will be generated after collect a certain amount of data. If we want to detect online data, data will be passed to SVM-predict module and predicted by BSVM algorithm will. At last system call Output module to output predicted results.

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. This study uses a famous port scanning software-Nmap, and DoS attack software-UDP flood to attack, uses SNORT to capture data. Because of network data contains a large number of invalid attributes or attributes insignificant to judge incursions, so we need to filter captured data first. By comparison and verification, we selected 9 most characteristic attributes *ip_proto, dport, th_flag, un_len, packet_flags, th_win, sport* to consist vector, as shown in Table 1.

This experiment adopts real attack data and online data for testing and verification.

Table.1 Results of derrenrent methods with some vector attributes

MultiPer	J48	muti-class bound-constrained SVM
94.2222	95.8272	93.2963
94.2716	95.8272	93.284
93.5432	95.5926	85.5679
95.0864	95.8272	93.284
91.5185	94.6543	90.1728
83.0617	95.7037	93.3333
93.9877	95.6914	94.9259
92.5309	94.3407	92.8519
82.1605	95.6914	94.9877

The most important part in SVM classification algorithm is model training. Model's quality determines accuracy and speed of classification. This study specializes trained and optimized the online data collected before and find the optimum model by contrast test.

Another essential factor in the SVM model is selection of kernel function. Because of existing research cannot give apposite kernel function selection method theoretically, this paper only selects it through a large number data from past experiments. The most suitable kernel function used for intrusion detection classification recognized in academia is radial basis function (RBF), so we use RBF as kernel function of SVM in this experiment.

RBF kernel function parameter (gamma, g) and the penalty coefficient (cost, c) are two other significant parameters of SVM model. This experiment applies famous K-fold cross-validation method to select these two parameters. Through divided training data into appropriate number of groups, train the first and second group then compares with the third group to get accuracy. And so on, until find a group with the highest accuracy then use script tests parameters g and c in the range of [c, g] = [2 ^ -10, 2 ^ 10] * [2 ^ -10, 2 ^ 10] one by one and select a group of g and c with highest accuracy. Finally, we obtain the best SVM model from existing training data.

This experiment utilizes original data set collected in previous step, through vectorization and normalization, conversed to data set that SVM can identify then generate SVM model. Test results of model kernel and kernel optimal are shown in Table 2 and Table 3, respectively.

Experimental results show that after attributes and model optimization, SVM-SNORT can reduce rate of False-negative and False-positives significantly in system. Due to small model, the detect speed reach nanosecond class, that will not occupy operating system's resources.

Table 2. Kernel function results

kerneltype	C	accuracy
linear	1	77.8519
polynomial		39.6296
radial basis function		93.2963
sigmoid		54.7513
linear	10	80.1975
polynomial		70.7531
radial basis function		93.4568
sigmoid		54.7531
linear	100	77.7531
polynomial		61.7654
radial basis function		93.5309
sigmoid		54.7531
linear	1000	80.9877
polynomial		59.5679
radial basis function		93.5062
sigmoid		54.7407

Table 3. Optimaling results

C	g	accuracy
1	0.3	95.0741
1	0.03	94.6049
1	0.003	94.1481
1	0.0003	82.4691
10	0.3	95.2469
10	0.03	94.9877
10	0.003	95.6049
10	0.0003	88.2099
100	0.3	95.3951
100	0.03	95.1235
100	0.003	95.4568
100	0.0003	90.4568
1000	0.3	95.3704
1000	0.03	95.284
1000	0.003	96.6543
1000	0.0003	91.2963

4. Conclusions

This paper proposes a novel SVM approach with K-fold Cross-Validation to optimize the attributes and SVM model in the IDS. These online data experimental results have been derived and analyzed. Compared with some representative machine learning method, online data experimental results show that this method can be used to reduce the rate of False-Negatives in the intrusion detection system.

5. Acknowledgment

Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality PHR 201106133.

6. References

[1] Dorothy E. Denning. “An intrusion detection model,” IEEE Transactions on Software Engineering, no. 2, pp. 222-232, 1987.

[2] P. Garcıa-Teodoro, J. Dıaz-Verdejo, G. Macia-Fernandez, and E. Vazquez. “Anomaly-based network intrusion detection: Techniques, systems and challenges,” Computers and Security, vol. 28, no. 1, pp. 18-28, 2009.

[3] T. F. Lunt, R. Jagannathan, R. Lee, A. Whitehurst, and S. Listgarten. "Knowledge-based intrusion detection," in Proceedings of the Annual Conference on AI Systems, 1989, pp.102-107.

[4] Li Shuwen, Artificial Invasion Detection Technique [J]. TAIYUAN SCI- TECH, no.02, 2006

[5] Vapnik V. The Nature of Statistical Learning Theory [M]. New York: Springer- Verlag, 1995

[6] LI Hui，GUAN Xiao-Hong，ZAN Xin ， and HAN Chong-Zhao, Network Intrusion Detection Based on Support Vector Machine[J], JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT, no.40, 2003

[7] Xiao Yun, Wang Xuanhong. Support Vector Machine theory and its application in network security. Xi'an Electronic and Science University Publishing House, 2011

[8] LIU Xin-zhe, Application of artificial intelligence techniques to Intrusion Detection System [J]. Railway Computer Application, no.08, 2004

[9] RAO Xian, DONG Chun-xi , YANG Shao-quan, Detecting intrusions by using support vector machines [J] JOURNAL OF XIDIAN UNIVERSITY, no.03, 2003

[10] YANG Ge, LI Yong—zhong, XU Jing，ZHAO BO. Network Intrusion Detection Based on Fuzzy Support Vector Classification Machines[J], Journal Of Jiangnan University(Natural Science Edition), no.06, 2007.