A Review on Intrusion Detection System Using Data Mining Technique: Support Vector Machine

Mistry Vaibhavi[1] and Patel Vibha[2]

Chhotubhai Gopalbhai Patel Institute of Technology, Bardoli, India
xyz, vibha.patel@utu.ac.in

Abstract. Intrusion Detection is the problem of identifying unauthorized use, misuse and abuse of computer systems. Outside attacks are not the only problem, the threat of authorized users misusing and abusing their privileges is an equally pressing concern. The proliferation of heterogeneous computer networks has additional implications for the intrusion detection problem. Namely, the increased connectivity of computer systems gives greater access to outsiders and makes it easier for intruders to cover their tracks. Therefore, the goals of an Intrusion Detection System (IDS) are to use all available information in order to detect both attacks by external hackers and misuse by insiders. IDS are based on the belief that an attacker's behaviour will be noticeable different from that of a legitimate user. Here, the focus is on Intrusion Detection System using data mining technique: SVM (Support Vector Machine). Classification is done by using SVM and comparative analysis of the system is done with neural network. Verification regarding the effectiveness of the system is done by conducting different experiments on different datasets such as, NSL-KDD Cup99 and DARPA.

Keywords: Classification, Intrusion Detection System (IDS), NSL- KDD, Support Vector Machine (SVM), Neural Network (NN).

1 Introduction

As network-based computer systems have important roles in modern society, they have become the targets of intruders. Therefore, we need to find the best possible ways to protect our systems. An intrusion can be defined as any action done to hamper the integrity, confidentiality or availability of the system. There are some intrusion prevention techniques which can be used to protect computer systems as a first line of defense. But only intrusion prevention is not enough. As systems become more complex, there are always exploitable weaknesses in the systems due to design and programming errors, or various penetration techniques. Therefore Intrusion detection is required as another measure to protect our computer systems from such type of vulnerabilities.

An Intrusion Detection System is used to detect all types of malicious network traffic and computer usage that can't be detected by a conventional firewall.

This includes network attacks against vulnerable services, data driven attacks on applications, host based attacks such as privilege escalation, unauthorized logins and access to sensitive files, and malware (viruses, Trojan horses, and worms).

There are several ways to categorize an IDS depending on the type and location of the sensors and the methodology used by the engine to generate alerts.

1.1 TYPES OF INTRUSION-DETECTION SYSTEMS

1. **Network Intrusion Detection System:** Identifies intrusions by examining network traffic and monitors multiple hosts. Network Intrusion Detection Systems gain access to network traffic by connecting to a hub, network switch configured for port mirroring, or network tap. An example of a NIDS is Snort.
2. **Host-based Intrusion Detection System:** Consists of an agent on a host which identifies intrusions by analyzing system calls, application logs file-system modifications and other host activities and state.
3. **Hybrid Intrusion Detection System:** Combines one or more approaches. Host agent data is combined with network information to form a comprehensive view of the network. An example of a Hybrid IDS is Prelude.

2 APPROACH FOR INTRUSION DETECTION SYSTEM

2.1 NEURAL NETWORK APPROACH FOR INTRUSION DETECTION SYSTEM

In order to apply this approach to Intrusion Detection, we would have to introduce data representing attacks and non-attacks to the Neural Network to adjust automatically coefficients of this Network during the training phase. In other words, it will be necessary to collect data representing normal and abnormal behavior to train the Neural Network. After training is accomplished, a certain number of performance tests with real network traffic and attacks were be conducted [1].

2.2 MACHINE LEARNING APPROACH FOR INTRUSION DETECTION

Intrusion Detection systems con not distinguish between normal and abnormal behavior of system using the audit data due to the ineffective behavior model system. Thus has to rely on the human for detection of the behavior. Involvement of Human in the detection system reduces the performances of the IDS as it become the tedious job with increasing data and network traffic. Due to the above deficiencies of IDSs based on human experts, intrusion detection techniques using machine learning have attracted more and more interests in recent

years. Machine learning is based heavily on statistical analysis of data and some algorithms can use patterns found in previous data to make decisions about new data [1].

Classification using Support Vector Machine: We have applied multi-class Support Vector Machines (SVMs) for classifier construction in IDSs and evaluated the performance of SVMs on the NSL-KDD dataset. The promising results clearly illustrate the learning efficiency and generalization ability of SVMs based on statistical learning theory[2]. Based on the idea of constructing optimal hyper-planes to improve generalization abilities, SVMs were originally proposed for binary classification problems. Nevertheless, most real world pattern recognition applications are multi-class classification cases. Thus, multi-class SVM algorithms have received much attention over the last decades and several decomposition-based approaches for multi-class problems have been proposed [3].

The idea of decomposition-based methods is to divide a multi-class problem into multiple binary problems, i.e., to construct multiple two-class SVM classifiers and combine their classification results. There are several strategies for the implementation of multi-class SVMs using binary SVM algorithms, which include one-vs.-all, one-vs.-one, and error correcting output coding (ECOC), etc. Among the existing decomposition approaches, the one-vs.-all strategy has been regarded as a simple method with relatively low precision when compared with other multi-class SVMs. However, a recent work in demonstrated that one-vs.-all classifiers are also extremely powerful and can produce results that are usually at least as accurate as other methods [4].

IDS is a valuable tool for the defense-in-depth of computer networks. Network-based IDS looks for known or potential malicious activities in network traffic and raise an alarm whenever a suspicious activity is detected. Several machine-learning techniques including neural networks, support vector machines (SVM), fuzzy logic have been studied for the design of IDS. In general, IDS deals with huge amount of data even for a small network, which contains irrelevant and redundant features. Extraneous features can make it harder to detect suspicious behavior patterns, causing slow training and testing process, higher resource consumption as well as poor detection rate. Feature selection is one of the key topics in IDS, it improves classification performance by searching for the subset of features, which best classifies the training data [5].

Reasons for using SVM: There are reasons that we use SVMs for intrusion detection. The first is speed: as real-time performance is of primary importance to intrusion detection systems, any classifier that can potentially run fast is worth considering. The second reason is scalability: SVMs are relatively insensitive to the number of data points and the classification complexity does not depend on the dimensionality of the feature space, so they can potentially learn a larger set of patterns and thus be able to scale better than neural networks. Once the data is classified into two classes, a suitable optimizing algorithm can be used if necessary for further feature identification, depending on the application.

3 RESULTS

3.1 RESULTS USING SVM

To verify the effectiveness and the feasibility of the proposed IDS system, we have used NSL-KDD dataset. It is a new version of KDDcup99 dataset. NSL-KDD dataset has some advantages over KDDcup99 dataset. It has solved some of the inherent problems of the KDDcup99, which is considered as standard benchmark for intrusion detection evaluation. The training dataset of NSL-KDD similar to KDDcup99 consist of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or attack type ,with exactly one specific attack type. The Intrusion Detection System is experimented using the Waikato Environment for Knowledge Analysis (WEKA 3.7).

The dataset is partitioned in to two classes: normal and anomaly, where the attack is the collection of all different attacks. The objective of our SVM experiments is to separate normal and anomaly patterns. In our case all attacks are classified as b, and normal data classified as a. we also apply SVMs to identify the most significant features for detecting attack patterns.

Table 1: Features of NSL-KDD dataset.

Sr no	Feature Name
1	Duration
2	Protocol_type
3	Service
4	Flag
5	Src_bytes
6	Dst_bytes
7	Land
8	Wrong_fragment
9	Urgent
10	Hot
11	Num_failed_logins
12	Logged_in
13	Num_compromised
14	Root_shell
15	Su_attempted
16	Num_root
17	Num_file_creations
18	Num_shells
19	Num_access_files
20	Num_outbound_cmds
21	Is_host_login
22	Is_guest_login
23	Count
24	Srv_count

25	Serror_rate
26	Srv_serror_rate
27	Rerror_rate
28	Srv_rerror_rate
29	Same_srv_rate
30	Diff_srv_rate
31	Srv_diff_host_rate
32	Dst_host_count
33	Dst_host_srv_count
34	Dst_host_same_srv_rate
35	Dst_host_diff_srv_rate
36	Dst_host_same_src_port_rate
37	Dst_host_srv_diff_host_rate
38	Dst_host_serror_rate
39	Dst_host_srv_serror_rate
40	Dst_host_rerror_rate
41	Dst_host_srv_rerror_rate
42	Normal or Attack

In the first set of experiment, the processed data consists of 22544 data points. We have composed training sets containing the same data points with, respectively, 42 features. Training is done using the Polykernel function; an important point of the kernel function is that it defines the feature space in which the training set examples will be classified. The results are summarized in the following table.

Table 2: SVM Training Summary

Correctly Classified Instances	21342	94.6682%
Incorrectly Classified Instances	1202	5.3318%
Time	-	103.02 Sec

Table 3: Confusion Matrix of SVM

Classified as	A	b
a=Normal	9000	711
b=Anomaly	491	12342

3.2 RESULTS USING NEURAL NETWORKS

For performance comparison with SVMs, the neural network experiments are considered to make binary normal/attack classification. In our experiments, we used a same dataset which we have used in SVM with the 2 classes of attack and anomaly and multi-layer, feed-forward network was trained.

Table 4: NN Training Summary

Correctly Classified Instances	21114	93.6568%
Incorrectly Classified Instances	1430	6.3432%
Time	-	121.19 Sec

Table 5: Confusion Matrix of NN

Classified as	a	B
a=Normal	9448	263
b=Anomaly	1167	11666

4 PERFORMANCE COMPARISON

Above results shows the performance of neural networks and support vector machines on the NSL-KDD data subset, using 42 features. We have taken 22544 instances for classification from which 21342 are correctly classified and 1202 are incorrectly classified. Accuracy of SVM is 94.66% and time taken is 103.02 seconds which is better compare to the result of NN. SVMs consistently outperform neural networks, in terms of training time and accuracy of detection. Even though the margin in accuracy is small and may not be statistically significant, there is an order of magnitude in the difference of training times. publication of the book:

We have conducted experiments using DARPA datasets also. The Cyber Systems and Technology Group (formerly the DARPA Intrusion Detection Evaluation Group) of MIT Lincoln Laboratory, under Defense Advanced Research Projects Agency (DARPA ITO) and Air Force Research Laboratory (AFRL/SNHS) sponsorship, has collected and distributed the first standard corpora for evaluation of computer network intrusion detection systems. The experimental results of NSL-KSS dataset and DARPA datasets are as follows.

Table 6: Comparison of SVM and NN

Dataset	Model	No of Instances	Correctly Classified	Incorrectly Classified	Accuracy	Time
NSL-KDD	SVM	22544	21342	1202	94.66%	103.02 sec
	NN	22544	21114	1430	93.65%	121.19 sec
DARPA Dataset2	SVM	500001	498556	1445	99.71%	95.59 sec
	NN	500001	499999	2	99.99%	655.41 sec
DARPA Dataset3	SVM	500001	499517	484	99.90%	15.09 sec
	NN	500001	499517	484	99.90%	27.9 sec
DARPA Dataset4	SVM	23183	23182	1	99.99%	0.13 sec
	NN	23183	23182	1	99.99%	29.6 sec

5 CONCLUSION

A number of experiments have been performed here, to measure the performance of support vector machines and neural networks in intrusion detection, using the NSL-KDD and DARPA datasets for intrusion evaluation. All classifications were performed on the binary (attack / anomaly) basis. Both SVMs and neural networks deliver highly-accurate performance, but SVMs perform better in terms of time. Even though SVMs are limited to making binary classifications, their superior properties of fast training, scalability and generalization capability give them an advantage in the intrusion detection application.

References

1. Deepika P Vinchurkar, Alpa Reshamwala, A Review of Intrusion Detection System Using Neural Network and Machine Learning Technique, International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 1, Issue 2, November 2012.
2. R.Rifkin, A.Klautau., In defense of one-vs.-all classification, Journal of Machine Learning Research, 5, pp.143-151, 2004.
3. T. G.Dietterich, G.Bakiri., Solving multiclass learning problems via error-correcting output codes, Journal of Artificial Intelligence Research, 2, pp. 263-286, 1995.
4. Yogita B. Bhavsar, Kalyani C.Waghmare, Intrusion Detection System Using Data Mining Technique: Support Vector Machine, International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013).

5. Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien, Ajith Abraham, Principle Components Analysis and Support Vector Machine based Intrusion Detection System.
6. Srinivas Mukkamala, Guadalupe Janoski, Andrew Sung, Intrusion Detection: Support Vector Machines and Neural Networks.
7. Jams Brentano, Steven R. Snapp, Gihan V. Dias, Terrance L. Goan, L. Todd Heberlein, Che-Lin Ho, Karl N. Levitt, Biswanath Mukherjee, Stephen E. Smaha, An Architecture for Distributed Intrusion Detection System.
8. J. Han., M. Kamber, J. Pei , Data Mining Concepts And Techniques, 3rd Edition, Morgan Kaufmann Publishers.

About Authors

Ms. Vaibhavi Mistry has completed her M.Tech in Computer Engineering from CGPIT in the year 2013 - 2015.

Vibha Patel is an Assistant Professor of Department of Computer Engineering and Information Technology, CGPIT at the Uka Tarsadia University of Gujarat, India. She received her M. Tech. in Computer Engineering from Dharmsinh Desai University. Her research interests include Data Mining and Data Analytics.