Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analysis on Intrusion Detection by Machine Learning Techniques: A Review

Manju Khari*, Anjali Karar
Ambedkar Institute of Advanced Communication Technologies & Research, New Delhi, India

Abstract: ***There are many risks of network attacks under the Internet environment, internet security is a vital issue and therefore, the intrusion detection is one major research problem for business and personal networks to resist external attacks. The goal of Intrusion detection systems (IDSs) is to provide a wall of defense to confront the attacks of computer systems on Internet where the conventional firewall do not succeeds. In this paper, comparison of most commonly used machine learning techniques based on genetic algorithm, support vector machines, artificial neural networks, etc. in intrusion detection domain, has been made. A comparative analysis of these techniques to detect intrusions in host and networks has also been made.***

Keywords: ***Intrusion detection system (IDS), anomaly detection, misuse or signature detection, genetic algorithm, machine learning technique, crossover, mutation***

I. Introduction

The Internet has become a part of daily life and an essential tool today. It aids people in many areas, such as business, entertainment and education, etc. In particular, Internet has been used as an important component of business models. For the business operation, both business and customers apply the Internet application such as website and e-mail on business activities. Therefore, security of using Internet as the media needs to be carefully concerned. Intrusion detection is one major research problem for business and personal networks.

As there are many risks of network attacks under the Internet environment, there are various systems designed to block the Internet-based attacks. Particularly, intrusion detection systems (IDSs) aid the network to resist external attacks. The goal of IDSs is to provide a wall of defense to confront the attacks of computer systems on Internet. IDSs can be used on detect difference types of malicious network communications and computer systems usage, whereas the conventional firewall cannot perform this task.

II. General Discussions On Intrusion Detection

In general, IDSs can be divided into two categories: Anomaly and misuse (signature) detection based on their detection approaches. Anomaly detection tries to determine whether deviation from the established normal usage patterns can be flagged as intrusions. On the other hand, misuse detection uses patterns of well-known attacks or weak spots of the system to identify intrusions. Genetic algorithm based intrusion detector (GBID) based on "learning the individual user behavior was proposed for intrusion detection, called. The user behavior is learnt by using genetic algorithms. Current user behavior can be predicted by genetic algorithm based on the past observed user behavior. The user behavior has been described using a 3-tuple (match index, entropy index, newness index). Value of the 3-tuple is calculated for fixed block size of commands in a user session called command sample. The 3-tuple value of a command sample in user session is compared with expected non-intrusive behavior 3-tuple value to find intrusions [1, 2]. The fuzzy logic has been used to detect false alarms. With fuzzy logic, the false alarm rate in determining intrusive activities can be reduced, where a set of fuzzy rules is used to define the normal and abnormal behavior in a computer network, and a fuzzy inference engine can be applied over such rules to determine intrusions. A genetic algorithm based technique has been proposed to generate fuzzy rules (instead of manual design) that are able to detect anomalies and some specific intrusions. Experiments were performed with DARPA data sets, which have information on computer networks, during normal behavior and intrusive behavior[3-5] . Genetic algorithm can be used to learn how to detect malicious intrusions and separate them from normal use. The algorithm has been tested in a real-world simulation to gauge its effectiveness under unpredictable conditions. Genetic algorithm method based on the theory of Darwinian evolution applied to mathematical models shows a high detection rate of malicious behavior and a low false positive rate of normal behavior classified as malicious. The genetic algorithm trained from data from which an empirical model of malicious computer behavior is generated, tested over previously unseen data to gauge its real-world performance, shows that the genetic algorithm is successfully able to generate an accurate empirical behavioral model from training data and then able to successfully apply this empirical knowledge to data never seen before. The final model

produced had an overall accuracy level of 97.8%, which showed both a high detection rate and an extremely low false positive rate. From these results, it could be concluded that genetic algorithms are a viable method for empirical model generation for computer intrusion detection. Genetic algorithms are now a possible alternative for the detection of malicious intrusions [6-9]. The pattern classification and knowledge discovery concept, which require the selection of a subset of attributes or features, cab be used to represent the patterns to be classified. Tasks such as medical diagnosis, a classification function learned through an inductive learning algorithm assigns a given input pattern to one of a finite set of classes. Typically, the representation of each input pattern consists of a vector of attribute, feature, or measurement values. The choice of features to represent the patterns affects several aspects of pattern classification [10]. Genetic Algorithm has been shown to improve support vector machines (SVM) based intrusion detection system (IDS). SVM is relatively a novel classification technique and has been shown higher performance than traditional learning methods in many applications. So several security researchers have proposed SVM based IDS. The fusion of GA and SVM enhances the overall performance of SVM based IDS. Through fusions of GA and SVM, the optimal detection model for SVM classifier can be determined. As the result of this fusion, SVM based IDS not only select "optimal parameters" for SVM but also "optimal feature set" among the whole feature set.[11, 12]

A misuse detection system based on genetic algorithm approach can be used for evolving and testing new rules for intrusion detection. To process network data in real time, principal component analysis (PCA) can be used to extract the most important features of the data. This results in keeping the high level of detection rates of attacks while speeding up the processing of the data. [13, 14]. The intrusion detection can be divided into three categories namely single, hybrid and ensemble. Single classifiers have been used extensively; however, ensemble classifiers outperform single classifiers in terms of classification accuracy. Due to recent developments in intrusion detection, it is now very difficult to design a single approach which outperforms the existing ones. The hybrid approaches have moved from marginalization to mainstream in the recent years. The hybrid approaches provide better flexibility and thus are gaining more popularity in the recent years [15].

III. Machine Learning Techniques

In the development of IDS, the ultimate goal is to achieve the best possible accuracy for the task at hand. This objective naturally leads to the design of some machine learning technique or to combine several machine learning techniques for the problem to be solved so that the system performance can be significantly improved. Based on the sources of the audit information used by each IDS, the IDSs may be classified into host based IDSs, distributed IDSs and network based IDSs. The Host based IDSs get audit data from host audit trails and it detect attacks against a single host. The distributed IDSs gather audit data from multiple hosts and possibly the network that connects the hosts and detect attacks involving multiple hosts. The network based IDSs use network traffic as the audit data source, relieving the burden on the hosts that usually provide normal computing services and detect attacks from network. In general, IDSs can also be divided into two categories based on their detection approaches: Anomaly detection and misuse detection. The anomaly detection determines whether deviation from the established normal usage patterns can be flagged as intrusions. It can detect any action that significantly deviates from the normal behavior. It is based on the normal behavior of a subject. Sometime assume the training audit data does not include intrusion data. Any action that significantly deviates from the normal behavior is considered intrusion. The misuse detection uses patterns of well-known attacks or weak spots of the system to identify intrusions. It can catch the intrusions in terms of the characteristics of known attacks or system vulnerabilities. It has advantages such as it is based on known attack actions, feature extract from known intrusions, integrate the Human knowledge and that the rules are pre-defined. The disadvantages include that it cannot detect novel or unknown attacks. While misuse detection can accurately and generate much fewer false alarms, it cannot detect novel or unknown attacks. Anomaly detection is able to detect unknown attacks based on audit, it suffers from disadvantage of having high false-alarm and limited by training data IDS are implemented by using some machine learning techniques which protect network and host from intruders before the attack has been launched. The final model produced must have higher accuracy levels, which shows both a high detection rate and an extremely low false positive rate. Some commonly used machine learning techniques are described below.

A. PATTERN CLASSIFICATION

Pattern recognition is the action to take raw data and activity on data category. The methods of supervised and unsupervised learning can be used to solve different pattern recognition problems. In supervised learning, it is based on using the training data to create a function, in which each of the training data contains a pair of the input vector and output (i.e. the class label). The learning (training) task is to compute the approximate distance between the input–output examples to create a classifier (model). When the model is created, it can classify unknown examples into a learned class labels.

B. SINGLE CLASSIFIERS

The intrusion detection problem can be approached by using one single machine learning algorithm. In literature, machine learning techniques (e.g. k-nearest neighbor, support vector machines, artificial neural network, decision trees, self-organizing maps, etc.) have been used to solve these problems.

C. HYBRID CLASSIFIERS

In the development of IDS, the ultimate goal is to achieve the best possible accuracy for the task at hand. This objective naturally leads to the design of hybrid approaches for the problem to be solved. The idea behind a hybrid classifier is to

combine several machine learning techniques so that the system performance can be significantly improved. More specifically, a hybrid approach typically consists of two functional components. The first one takes raw data as input and generates intermediate results. The second one will then take the intermediate results as the input and produce the final results. In particular, hybrid classifiers can be based on cascading different classifiers, such as neuro-fuzzy techniques. On the other hand, hybrid classifiers can use some clustering-based approach to preprocess the input samples in order to eliminate unrepresentative training examples from each class. Then, the clustering results are used as training examples for classifier design. Therefore, the first level of hybrid classifiers can be based on either supervised or unsupervised learning techniques. Finally, hybrid classifiers can also be based on the integration of two different techniques in which the first one aims at optimizing the learning performance (i.e. parameter tuning) of the second model for prediction.

D. ENSEMBLE CLASSIFIERS

Ensemble classifiers were proposed to improve the classification performance of a single classifier. The term ‘‘ensemble” refers to the combination of multiple weak learning algorithms or weak learners. The weak learners are trained on different training samples so that the overall performance can be effectively improved. Among the strategies for combining weak learners, the ‘‘majority vote” is arguably the most commonly used one in the literature. Other combination methods, such as boosting and bagging, are based on training data re-sampling and then taking a majority vote of the resulting weak learners.

IV. Conclusion

From the comparative analysis on the various machine learning techniques for the intrusion detection, it is concluded that the genetic algorithms are a viable method for the detection of malicious intrusions. It is also established that the probabilities of crossover, mutation, and the selection of the highest ranked individual in some hybrid classifiers are close to results achieved by using genetic algorithm. The comparison between various learning techniques will allow software professionals to find best machine learning technique to find clear, unambiguous knowledge about intrusion detection more effectively and efficiently. It can also assist in making acceptable tradeoffs among sometimes conflicting goals such as functionality, quality, cost, and time to market and to allocate resources based on the security requirements importance to the project as a whole.

References

[1] ShengYi Jiang, Xiaoyu Song , Hui Wang, Jian-Jun Han, Qing-Hua Lid,,A clustering-based method for unsupervised intrusion detections, School of Informatics, GuangDong University of Foreign Studies, Guangzhou, Guangdong, China, Electrical and Computer Engineering, Portland State University, Oregon, OR, USA, Communication Command College, Wuhan, Hubei, China, Computer School, Huazhong University of Science and Technology, Wuhan, Hubei, China, Pattern Recognition Letters, Elsevier 27 (2006) 802–810

[2] H. Gunes Kayacik, A. Nur Zincir-Heywood, Malcolm ,A hierarchical SOM-based intrusion detection system I. Heywood Dalhousie University, Faculty of Computer Science, University Avenue, Halifax, NS, Canada, Engineering Applications of Artificial Intelligence, Elsevier 20 (2007) 439–451

[3] Latifur Khan, Mamoun Awad, Bhavani Thuraisingham, A new intrusion detection system using support vector machines and hierarchical clustering, Springer-Verlag, Springer-Verlag 2006

[4] Ozgur Depren, Murat Topallar, Emin Anarim, M. Kemal Ciliz, An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks, Bogazici University, Electrical and Electronics Engineering Department, Information and Communications Security (BUICS) Lab, Bebek, Istanbul, Turkey, Expert Systems with Applications 29 (2005) 713–722

[5] Sampada Chavan, Khusbu Shah, Neha Dave and Sanghamitra Mukherjee,Adaptive Neuro-Fuzzy Intrusion Detection Systems, Institute of Technology for Women, SNDT University, India, Ajith Abraham, Department of Computer Science, Oklahoma State University, USA, Sugata Sanyal, School of Tech. and Computer Science, Tata Institute of Fundamental Research, India, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC’04) 0-7695-2108-8/04 © 2004 IEEE

[6] Mohammad Saniee Abadeh, Jafar Habibi, Zeynab Barzegar, Muna Sergi,A parallel genetic local search algorithm for intrusion detection in computer networks, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran, Engineering Applications of Artificial Intelligence, Elsevier, 20 (2007) 1058–1069

[7] Giorgio Giacinto and Fabio Roli, Intrusion Detection in Computer Networks by Multiple Classifier Systems, Department of Electrical and Electronic Engineering - University of Cagliari, Italy, Piazza D - 09/23 Cagliari, Italy, Jgiacinto,roli, iee.unica.it, 2002, IEEE.

[8] German Florez, Susan M. Bridges, and Rayford B. Vaughn, ,An Improved Algorithm for Fuzzy Data Mining for Intrusion Detection, 0-7803-7461-4/02/ 2002 IEEE

[9] Jonatan Gomez and Dipankar ,Evolving Fuzzy Classifiers for Intrusion Detection Dasgupta, Proceedings of the 2002 IEEE, Workshop on Information Assurance, United States Military Academy, West Point, NY June 2001

[10] Latifur Khan , Mamoun Awad , Bhavani Thuraisingham, A New Intrusion Detection using support vector machines and hierarchical clustering, Springer-Verlag 2006

[11] Mohammad Sazzadul Hoque, Md. Abdul Mukit and Md. Abu Naser Bikas, An implementation of Intrusion Detection system using genetic algorithm, International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012

[12] Dong Seong Kim, Ha-Nam Nguyen, Jong Sou Park,Dept. of Computer Engineering, Hankuk Aviation University, Seoul, Korea, Genetic Algorithm to Improve SVM Based Network Intrusion Detection System

[13] Zorana Bankovic, Dus an Stepanovic, Slobodan Bojanic, Octavio Nieto-Taladriz, Improving network security using genetic algorithm approach, ETSI Telecomunicacio´ n, Technical University of Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain, Faculty of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 78, 11000 Beograd, Serbia

[14] B.balajinath and S.V.Raghavan, Intrusion detection through learning behavior model, network and Multimedia systems labortary, Department of computer science engineering, Indian Institute of Technology Madaras, 13 november 2000

[15] Chih-Fong Tsai , Yu-Feng Hsu , Chia-Ying Lin , Wei-Yang Lin, Intrusion detection by machine learning: A Review, Expert Systems with Applications 36 (2009) 11994–12000

