ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

Data Mining-Based Intrusion Detection Systems

B.Muthulakshmi[1], Dr.V.Thiagarasu[2]

[1]M.Phil Research Scholar, PG & Research, Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam, India.

[2]Associate Professor, PG & Research, Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam, India.

ABSTRACT: Intrusion Detection Systems are designed to detect system attacks and it classifies system activities into normal and abnormal form. Data Mining based intrusion detection system model generalizes and detects both known attacks and normal behaviour in order to detect unknown attacks and fails to generalize and detect new attack without known signatures. Intrusion detection system faces three types of issues such as accuracy, efficiency and usability. Intrusion Detection System is used to detect all kinds of new attacks which can be implemented using machine learning techniques with high accuracy. The machine learning techniques are Decision trees, K-nearest neighbour, Ripper rule, Bayesian Network and they are used to analyzed and reducing the false alarm rate. This work compares the accuracy between two intrusion detection systems to examine their Receiver Operating Characteristics (ROC) curves.

KEYWORDS: Intrusion Detection, Receiver Operating Characteristics, Back-Propagation Neural Network, False Alarm Rate, Machine Learning.

I. INTRODUCTION

Security of network systems is becoming increasingly important as more and more sensitive information is being stored and manipulated online. Intrusion Detection Systems (IDSs) have thus become a critical technology to help protect these systems. Most IDSs are based on hand-crafted signatures that are developed by manual encoding of expert knowledge. These systems match activity on the system being monitored to known signatures of attacks. The problem with this approach is that these IDSs fail to generalize to detect new attacks or attacks without known signatures. There is an increasing interest in data mining based approaches to building detection models for IDSs. These models generalize from both known attacks and normal behaviour in order to detect unknown attacks. IDS can also be generated in a quicker and more automated method than manually encoded models that require difficult analysis of audit data by domain experts.IDS better detection performance and generalization ability of data mining based IDSs are some difficulties in the implementation of the system [1]. This prevents them from being able to process audit data and detect intrusions in real time. An effective intrusion detection system should work in real-time, as intrusions take place, to minimize security negotiation [2]. Elimination of the insignificant and useless inputs leads to a simplification of the problem, and faster and more accurate detection results [3]. The proposed approach of IDS is comparing with the accuracy, efficiency and usability using the machine learning techniques with the detecting system.

II. RELATED WORK

IDS work most similar to unsupervised model generation is a technique developed at SRI in the Emerald system [4]. Emerald uses historical records to build normal detection models and compares distributions of new instances to historical distributions. Inconsistency between the distributions signifies an intrusion. Related to automatic model generation is adaptive intrusion detection. Teng et al. [5] perform adaptive real time anomaly detection by using inductively generated sequential patterns. Sobriety's work on adaptive intrusion detection using an expert system to

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

collect data from audit sources [6]. Hellman and Bhangoo [7] present a statistical method to determine sequences which occur more frequently in intrusion data as opposed to normal data. Lee et al. [8] uses a prediction model trained by a decision tree applied over the normal data. Lane and Bradley [9] examine unlabeled data for anomaly detection by looking at user profiles and comparing the activity during an intrusion to the activity under normal use.

III. PROPOSED SYSTEM

Intrusion detection system planned the machine learning deals with the task of building programs that improve their performance through experience. Machine learning algorithms have proven to be of great value in a variety of domains [10]. They are particularly useful for (a) poorly understood problem domains where little knowledge exists for the humans to develop effective algorithms; (b) domains where there are large databases containing valuable hidden regularities to be discovered; or (c) domains where programs must adapt to changing conditions. Machine learning is basically used to give the high detection accuracy for real-time intrusion detection system and improvement of algorithms that are existing for the same. Basically, input given to machine learning technique is empirical data and output of this technique is the patterns/features of the underlying mechanism that generated the data. Many approaches for machine learning techniques are:

- Decision tree
- Ripper Rule
- Back-Propagation Neural Network
- Bayesian Network
- Radial Basis Function Neural Network.

Decision tree

Decision tree concept is basically used in data mining to efficiently classify the data. It consists of non-terminal and terminal nodes. Non terminal means a root and internal node, and the terminal nodes are (leaves). Initially, decision trees classify the known data and untrained data in decision tree by identifying attribute and value that will be used input data at each internal node after training. The data decision tree traverse from the starting of root node to internal node [17]. Algorithms for decision tree are (a) C4.5 Algorithm and (b) ID3 Algorithm.

Ripper Rule

Ripper is the (Repeated Incremental Pruning to Produce Error Reduction) efficient rule based learning algorithm process the various noisy dataset. It should be noted that ripper is used to handle data sets with the target that take on more than two unique values. It consist of two stages (a) first initialize the rule condition. (b) Usages rule optimization technique [18]. Consisting of two loops:

- Outer loop
- Inner loop

Bayesian Network

A Bayesian Network involves both a graphical model and a probabilistic model representing random variables and condition dependence through a DAG . Nodes in the graph represent random variables, edges represent conditional dependencies. Thus nodes that are not connected represent variables that are conditionally independent on each other [20].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

Back-Propagation Neural Network

Back-Propagation Neural Network (BPNN) model is a well-known supervised learning model that can effectively classify many types of data. BPNN is a feed-forward multi-layer network. Its input vectors and the corresponding target vectors are used in training the network until it can approximate a function [19].

A. ACCURACY

To design and apply the effective data mining-based IDS is defining specifically how detection accuracy of these systems is measured. The difference in nature between a data mining-based system and a typical ID, the evaluation metrics must take into account factors which are not important for traditional IDSs. There are several key component is used to measure the accuracy. The key components are detection rate, which is the percentage of attacks that a system detects. Another component is the false positive rate, which is the percentage of normal data that the system falsely determines to be intrusive. A method is to compare accuracy between two IDSs is to check their ROC curves. Data mining-based systems have the advantage of potentially being able to detect new attacks that hand-crafted methods are used for detection rate is higher than false positive rate [11].

B. EFFICIENCY

Effective intrusion detection should happen in real-time, as intrusions take place, to minimize security compromises. The data mining-based ID models work efficiently for real-time intrusion detection. In contrast to off-line IDSs, a key objective of real-time IDS is to detect intrusions as early as possible. DoS attacks, which typically generate a large amount of traffic in a very short period time, are often used by intruders to first overload an IDS, and use the detection delay as a window of opportunity to quickly perform their malicious intent. It is necessary to examine the time delay associated with computing feature in order to speed up model evaluation [12]. The time delay of a feature includes not only the time spent for its computation, but also the time spent waiting for its readiness.

C. USABILITY

A data mining-based IDS are significantly more complex than a traditional system. The main cause for this is that data mining systems require large sets of data from which to prepare. The expectation to reduce the complexity of data mining systems has led to many active research areas [13, 14]. First, management of both training and historical data sets is a difficult task, especially if the system handles many different kinds of data. Second, once new data has been analyzed, models need to be updated. Third, many data mining-based IDSs are difficult to deploy because they need a large set of clean labelled training data. Typically the attacks within the data must either be manually labelled for training signature detection models, or removed for training anomaly detection models [15].

D. REDUCING FALSE ALARMS

The system is still giving some false alarms for all the four algorithms some more training is needed to be given. This is the machine learning mechanism i.e. the system will keep on learning on its own without human interference and there is no updating required. The alternative way to using a majority voting algorithm for every two consecutive detection results for each pair of source and destination IP addresses. To verify the result is normal network activity or is an attack type by grouping the network data from the classification phase into groups of two records. In each group, if there are at least 3 or 4 records which are reported to be the same attack type, then this group of data is considered the attack and also the data is considered as normal. This procedure can increase the detection accuracy and the user's confidence in the alarms provided IDS.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

IV. RESULT AND DISCUSSION

The result of intrusion detection system is to compare the ROC curve of two systems accuracy is low. Intrusion detection system compares the result in before detection and after detection with accuracy, efficiency, usability of low, medium and high. The result of algorithms will then be scrutinized and techniques for reducing false alarm rate and increasing the accuracy will be tried and tested. Table.1 shows before detection with intrusion detection system. Table.2 shows after intrusion detection system with accuracy is high compare with efficiency and usability [16].

Table 1: Before detection

	System1			System2		
	Low	Medium	High	Low	Medium	High
Accuracy	No	Yes	No	No	No	Yes
Efficiency	Yes	No	No	Yes	No	No
Usability	No	Yes	No	No	Yes	No

Table 2: After detection

	System1			System2		
	Low	Medium	High	Low	Medium	High
Accuracy	No	No	Yes	No	No	Yes
Efficiency	Yes	No	No	Yes	No	No
Usability	No	Yes	No	No	Yes	No

V. CONCLUSION AND FUTURE WORK

In this paper, outlined the breadth of research efforts to address important and challenging issues of accuracy, efficiency, and usability of real-time IDSs. To implement the feature extraction and construction algorithms for labelled audit data. So far, there have been very few or no tries at attempting to develop a real-time IDS. All prior systems focus on Offline traffic only. To examined various algorithms like Decision Tree, k-NN, BPNN and Ripper Rule and will try to implement them in practice. Hence, using best algorithms can implement real time online network intrusion detection system. The research efforts on IDSs for e-Commerce and e-Government applications in the near future. This research will be to take traffic that comes directly from the network and try to prevent the intrusions by making the False Alarm Rate low or nil.

REFERENCES

[1] E. Eskin. Anomaly detection over noisy data using learned probability distributions 2009.
[2] A. Ghosh and A. Schwartzbard. A study in using neural networks for anomaly and misuse detection. 2009.
[3] W. Lee, S. J. Stolfo, and K. Mok. Data mining in work flow environments: Experiences in intrusion detection. 2010
[4] W. Lee, R. Nimbalkar, K. Yee, S. Patil, P. Desai, and Emerald. A data mining and CIDF based approach for detecting novel and distributed intrusions 2008.
[5] Teng et al. Intrusion detection with unlabeled data using clustering. Columbia University,2009.
[6] M. Sobirey, B. Richter, and M. Konig. The intrusion detection system aid. architecture, and experiences in automated audit analysis 2006.
[7] S. Stainford-Chen. Common intrusion detection framework.
[8] S. Staniford-Chen, B. Tung, and D. Schnackenberg. The common intrusion detection framework (cidf) 2007.
[9] Lane and Bradley Cost-sensitive modeling for fraud and intrusion detection 2008.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

[10] S.X. Wu, W. Banzhaf, The use of computational intelligence in intrusion detection system: a review, Applied Soft Computing 10 (2010) 1–35.
[11] H. S. Teng, K. Chen, and S. C. Lu. Adaptive real-time anomaly detection using inductively generated sequential patterns 2008.
[12] M. Amini, A. Jalili, H. Reza Shahriari, RT-UNNID: a practical solution to realtime network-based intrusion detection using unsupervised neural networks, Computer & Security 25 (2005) 459–468.
[13] N. Ngamwitthayanon, N. Wattanapongsakorn, C. Charnsripinyo, D.W.Coit, Multi-stage network-based intrusion detection system using back propagation neural networks, in: Asian International Workshop on Advanced Reliability Modeling (AIWARM), Taiwan, 2008, pp. 609–619.
[14] M-Y. Su, G-J. Yu, C-Y. Lin, A real-time network intrusion detection system for large-scale attacks based on an incremental mining approach, Computers and Security 28 (2009) 301–309.
[15] Z. Li, Y. Gao, Y. Chen, HiFIND: a high-speed flow-level intrusion detection approach with DoS resiliency, Computer Networks 54 (2010) 1282–1299.
[16] W. Lee, S. J. Stolfo, and K. Mok. Data mining in work flow environments: Experiences in intrusion detection 2011.
[17] M. Panda, M.R. Patra, Semi-Naive Bayesian method for network intrusion detection system, Neural information processing, Lecture Notes in Computer Science (Springer Link) 5863 (2012) 614–621.
[18] N. Ngamwitthayanon, N. Wattanapongsakorn, C. Charnsripinyo, D.W.Coit, Multi-stage network-based intrusion detection system using back propagation neural networks, Taiwan, 2012, pp. 609–619.
[19] P. Sangkatsanee, N. Wattanapongsakorn, C. Charnsripinyo, Network intrusion detection with artificial neural network, decision tree and rule based approaches, 2012.
[20] R.S. Puttini, Z. Marrakchi, L. Me, A Bayesian classification model for real-time intrusion detection, vol. 659, 2011, pp. 150–162.