Expert Systems With Applications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

An intrusion detection system using network traffic profiling and online sequential extreme learning machine

Q1 Raman Singh [a,*], Harish Kumar [b], R.K. Singla [c]

[a] University Institute of Engineering and Technology, Panjab University (UIET), Chandigarh, (India)
[b] University Institute of Engineering and Technology (UIET), Panjab University, Chandigarh, (India)
[c] Department of Computer Science and Applications (DCSA), Panjab University, Chandigarh, (India)

ARTICLE INFO

Keywords:
Intrusion detection system
Feature selection technique
Network traffic dataset
Network traffic profiling
Online sequential extreme learning machine (OS-ELM)

ABSTRACT

Anomaly based Intrusion Detection Systems (IDS) learn normal and anomalous behavior by analyzing network traffic in various benchmark datasets. Common challenges for IDSs are large amounts of data to process, low detection rates and high rates of false alarms. In this paper, a technique based on the Online Sequential Extreme Learning Machine (OS-ELM) is presented for intrusion detection. The proposed technique uses alpha profiling to reduce the time complexity while irrelevant features are discarded using an ensemble of Filtered, Correlation and Consistency based feature selection techniques. Instead of sampling, beta profiling is used to reduce the size of the training dataset. For performance evaluation of proposed technique the standard NSL-KDD 2009 (Network Security Laboratory-Knowledge Discovery and Data Mining) dataset is used. In this paper time and space complexity of the proposed technique is also discussed. The experimental results yielded an accuracy of 98.66% with a false positive rate of 1.74% and a detection time of 2.43 s for binary class NSL-KDD dataset. The proposed IDS achieve 97.67% of accuracy with 1.74% of false positive rate in 2.65 s of detection time for multi-class NSL-KDD dataset. The Kyoto University benchmark dataset is also used to test the proposed IDS. Accuracy of 96.37% with false positive rate of 5.76% is yielded by the proposed technique. The proposed technique outperforms other published techniques in terms of accuracy, false positive rate and detection time. Based on the experimental results achieved, we conclude that the proposed technique is an efficient method for network intrusion detection.

1. Introduction

In the world of rapidly developing technology, networks are facing threats like viruses, worms, Trojan horses, spyware, adware, root kits, etc. These intrusions need to be identified before any type of loss to the organizations. Even internal Local Area Network (LAN) is also seriously struggling with intrusions. This is affecting productivity of computer networks in terms of bandwidth and other resources. Hackers use advance features like dynamic ports, IP address spoofing, encrypted payload etc., to avoid detection. This type of intrusions can be detected by discovering patterns in network traffic dataset. Due to huge and imbalanced dataset machine learning based Intrusion Detection System (IDS) faces problem to process entire data. So, it is necessary to identify intrusions through network traffic behavior.

IDS is designed to defend the network from malicious activities. Anomaly based IDS learn normal behavior from network traffic dataset to detect attacks. Soft computing based IDS embraces several computational intelligence methodologies, including artificial neural networks, fuzzy logic, evolutionary computation, probabilistic computing, artificial immune systems, belief networks etc.

This paper presents an intrusion detection technique that considers various issues like hugeness of network traffic dataset, feature selection, low accuracy and high rate of false alarms. Online Sequential Extreme Learning Machine (OS-ELM) (Liang, Huang, Saratchandran, & Sundararajan, 2006) is used to process network traffic dataset to detect intrusions. It is fast and accurate single hidden layer feed forward neural network (SHLFN) which can process network instances one by one or in chunks. It has proved its applicability in classification by performing in single iteration. The performance evaluation of proposed technique is carried out using benchmarked NSL-KDD 2009 (Network Security Laboratory-Knowledge Discovery and Data Mining) dataset (Tavallaee, Bagheri, Lu, & Ghorbani, 2009) and Kyoto University Benchmark dataset (Kyoto, 2009). The experimental results show that proposed IDS is able to achieve higher accuracy with low false alarm rates. This approach is able to address high dimensionality issue of network traffic dataset.

The rest of the paper is organized as follows: Section 2 confers the related literature about intrusion detection, dimensionality reduction

* Corresponding author. Tel.: +919530802235.
E-mail addresses: raman.singh@ieee.org (R. Singh), harishk@pu.ac.in (H. Kumar), rksingla@pu.ac.in (R.K. Singla).

http://dx.doi.org/10.1016/j.eswa.2015.07.015
0957-4174/

and online sequential extreme learning machine. Section 3 discusses network traffic dataset used. This section also presents the proposed IDS methodology. Section 4 examines the results. Finally, Section 5 concludes the study along with future directions.

2. Related work

This section presents the related work carried out by various researchers.

2.1. Intrusion detection techniques

IDS can be classified by signature, anomaly or hybrid technique. There are many soft computing techniques which have been used to detect intrusions. These techniques along with various issues are discussed here. Anomalies are detected by analyzing sparse region of this feature space. Genetic clustering based intrusion detection automatically creates normal and abnormal clusters and effectively detects intrusions. This unsupervised technique operates in two phases. In first phase network data is grouped by taking nearest neighbor. Second phase obtains near optimal detection rate by genetic optimization (Liu, Chen, Liao, & Zhang, 2004). Artificial Neural Network (ANN) and Support Vector Machine (SVM) are also used in IDS. Two encoding methods simple frequency-based and term frequency × inverse document frequency (*tf* × *idf)* scheme are used to detect possible intrusions (Chen, Hsu, & Shen, 2005). Patterns of intrusions are built by IDS using random forest of training instances. After learning these patterns, intrusions are detected by the outlier detection algorithm. This hybrid approach improves the detection rate of IDS (Zhang, Zulkernine, & Haque, 2008). AdaBoost can be used to increase the efficiency of IDS. In this technique decision stump is used as a weak classifier. The decision rules are defined for both categorical and continuous features. Both types of features are combined to form a strong classifier (Hu, W. & Maybank, S., 2008).

Concept drifting data streams is used in adaptive ensemble approach for classification. This model is updated automatically by traditional mining classification. Three classification algorithms namely Expectation–Maximization (EM), C4.5 and K-Nearest Neighbor (KNN) are employed to test performance of the system (Farid, Zhang, Hossain, Rahman, Strachan, Sexton, & Dahal, 2013). Hybrid Intelligent Intrusion Detection and Prevention System (IIDPS) is used to detect intrusions in early stage. Known attacks are identified by the signature based approach but others are detected using anomaly based approach. The support vector machine (SVM) with three types of kernel (Linear, polynomial and RBF) is used to detect unknown attacks in (Alazab, Hobbs, Abawajy, Khraisat, & Alazab, (2014)). IDS faces problem to process huge network traffic dataset. Due to this learning is slow and detection time of intrusions increases. The dataset is also highly imbalanced. The difficulty to clearly separate normal and anomalous behavior decreases accuracy and increase false positive rate (Elshoush, 2011).

From the study of various techniques it has been found that anomaly detection techniques has a low accuracy rate. Anomaly based IDS also suffer with a high false alarm rate. These issues are research challenges in the field of anomaly based IDS. There is need of IDS which deals with issues like hugeness of network traffic dataset, low accuracy and high false alarm rate. The hugeness of network traffic dataset can be addressed by feature reduction.

Feature reduction retains high quality features but discarding redundant and irrelevant features. Feature selection techniques are used to reduce the number of features of network traffic dataset. A new feature selection approach based on cuttlefish optimization algorithm is used for IDS. Researchers suggest that feature selection should be used to remove non-useful features. Authors also prove that accuracy can be increased while false positive can be decreased by selecting good quality of features (Eesa, Orman, & Brifcani, 2015). It is realizing from the literature that the irrelevant features should be removed even if the dataset contains few features. It is always best practice to remove irrelevant feature and select good quality features. This reduces memory space requirement and detection time. Redundant and duplicate instances can affect the performance of machine learning techniques. Due to these replicated instances, these techniques may become biased. This is also the reason for over fitting. So preprocessing of dataset is required to discard these redundant and duplicate instances. Due to the computational power and memory limitations of IDS, it is still difficult to process whole dataset. Sampling can be used to reduce the training dataset but researchers suggest that due to imbalances in network traffic dataset there may be loss of information. So, sampling is not suggested for network traffic dataset (Singh, Kumar, & Singla, 2013).

Sampling and feature selection techniques may improve the analysis of huge data set. But these operations may change overall characteristics of data. Alpha and beta profiling can reduce the effect of imbalanced dataset. Alpha profiling ensures that each type of instance remains in training dataset. Beta profile can reduce the sample size of training dataset while keeping characteristics of dataset intact. From the literature it is found that alpha profiling has not been used to reduce the effect of imbalances and detection time. It is also found that beta profiling has not been used efficiently to reduce the size of training sample.

2.2. Online sequential extreme learning machine (OS-ELM)

OS-ELM is designed to overcome the slow learning limitation of feed forward neural network. It provides good generalization performance with fast learning speed (Liang et al., 2006). Fuzzy OS-ELM solves the approximation and classification problem. Target echo signals of high resolution range radars are used in the target recognition system. The extreme learning machine is used to increase the learning speed of the system (Avci, 2012). Non-stationary time series is predicted by the online sequential extreme learning machine with kernels (OS-ELMK). A limited memory prediction strategy provides good accuracy with at least an order-of-magnitude reduction in the learning time (Wang, 2014). Role of links and node features in social network is identified by the OS-ELM based node classification technique. Node features as well as interaction between nodes are considered for classification (Sun, Yuan, & Wang, 2015). Extreme Learning Machine (ELM) is also used to detect intrusions in IDS. Researchers also compared performances of SVM and ELM. It is found that SVM and ELM provide comparable accuracy but ELM performs faster than SVM. ELM takes less time than SVM in order to detect intrusions. Very small sample set are used and there is need of performance evaluation of OS-ELM for larger network traffic dataset (Cheng, Tay, & Huang, 2012).

From the literature, it has been found that OS-ELM is an emerging classification technique. This technique has been used to solve many classification problems. It can process large dataset in less time which makes it good candidate for IDS. The literature encourages the use of OS-ELM to detect intrusions in computer networks.

3. Network traffic dataset and methodology

This section describes the network traffic dataset used for performance evaluation of proposed IDS. It also explains various experiments performed on this dataset in order to detect intrusions.

3.1. Network traffic dataset

The Cyber Systems and Technology Group of MIT Lincoln Laboratory have collected network traffic dataset. This lab simulated U.S. Air Force LAN with multiple attacks and acquire nine week of TCP dump

Please cite this article as: R. Singh et al., An intrusion detection system using network traffic profiling and online sequential extreme learning machine, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.015