Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 1–6

Procedia Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

Performance evaluation of intrusion detection based on machine learning using Apache Spark

Mustapha Belouch[a,*], Salah El Hadaj[a], Mohamed Idhammad[b]

[a]*Faculty of Science and Technics, Cadi Ayyad University, Marrakech, Morocco*
[b]*Faculty of Science, Ibn Zohr University, Agadir, Morocco*

Abstract

Nowadays, network intrusion is considered as one of the major concerns in network communications. Thus, the developed network intrusion detection systems aim to identify attacks or malicious activities in a network environment. Various methods have been already proposed for finding an effective and efficient solution to detect and prevent intrusion in the network, ensuring network security and privacy. Machine learning is an effective analysis framework to detect any anomalous events occurred in the network traffic flow. Based on this framework, the paper in hand evaluates the performance of four well-known classification algorithms; SVM, Naïve Bayes, Decision Tree and Random Forest using Apache Spark, a big data processing tool for intrusion detection in network traffic. The overall performance comparison is evaluated in terms of detection accuracy, building time and prediction time. Experimental results on UNSW-NB15, a recent public dataset for network intrusion detection, show an important advantage for Random Forest classifier among other well-known classifiers in terms of detection accuracy and prediction time, using the complete dataset with all 42 features.

Keywords: Intrusion Detection; Machine Learning; Apache Spark.

1. Introduction

Intrusion is defined as a set of activities that violates security objectives. In general, an intrusion detection system (IDS) must analyze the network traffic and immediately warn of potential threats. When any malicious intrusion or attack is manifest in a network, computers and information systems may suffer serious consequences when defined computer security policies are violated. Various security strategies have been employed over the years to safeguard networks. The firewall is considered as a basic packet filter; however, it has been proved that is not sufficient in providing a secure network environment. Intrusion detection working in conjunction with a firewall may provide an improved and safer network. In the literature, many IDSs have been developed via the implementation of various techniques derived from different disciplines including statistical methods, AI techniques, and more. Some IDSs are based on single-classification techniques, while others (hybrid/ensemble) include more than one classification techniques. Ensemble based IDSs present many advantages over the single classification IDS. Many works have

* Corresponding author. Tel.: +212672220196.
E-mail address: mbelouch@gmail.com (M. Belouch)., elhadajs@yahoo.fr (S. Elhadaj)., idhammad.mohamed@gmail.com (M. Idhammad).

10.1016/j.procs.2018.01.091

proposed different ensembles for ID, primarily through the exploitation of different characteristics of weak classifiers and datasets.

IDS are categorized by misuse and anomaly strategies. The misuse approach seeks known attacks referred to as attack signatures, while the anomaly approach is based on normalcy models, where any significant deviation from these reference models indicates a potential threat. However, both approaches suffer from a number of weaknesses. Misuse detection requires frequent updates of signatures to ensure ample detection, while anomaly detection presents a high false positive rate. Thus, the current challenge is to tackle these two shortcomings toward the provision of a solution with characteristics such as superior accuracy with low false positive rates.

The solution of using multiple classifiers has been widely employed to solve various classification problems, including IDS [1, 2, 3]. These methods are based on feature representation, with an accurate voting system and weighting assignment that can improve the classification rate. Furthermore, in domains with huge data volumes, such as network traffic, available resources and computational time are greatly impacted.

The purpose of this paper is to address the issue of low accuracy and prediction time in IDS. We employed multiple classifiers with different learning paradigms to evaluate different classifier model. The organization of this paper is as follows: The majority of Section 2 discusses the background and related works. Section 3 presents the various classification techniques and dataset employed in this work, and Section 4 provides experimental results and a discussion on the findings. Finally, Section 5 concludes the paper.

2. Related Work

Several research papers describing the use of machine learning methods for anomaly detection have reported the attainment of a very high detection rate of 98%, while the false positive rate was less than 1% [4]. However, when surveying state-of-art IDS solutions and commercial tools, there is no evidence for the utilization of anomaly detection approaches, presumably because experts still consider anomaly detection to be an immature technology. To discover the reason for this contrast [5] concentrated studies on the details of the research done in anomaly detection were considered from different angles, for example learning and detection approaches, training data sets, testing data sets, and evaluation methods. These studies indicated that there are some issues in the KDDCUP'99 data set [6], which is broadly used as one of the rare publicly accessible data sets for network based anomaly detection systems.

The primary critical deficiency in the KDD data set was the enormous number of redundant records. Approximately 78% and 75% of the train and test set records, respectively, are duplicated in KDD [5]. These redundant records in the train set cause learning algorithms to be one sided toward the more frequent records, and prevent it from learning infrequent records, which are typically more damaging. On the other hand, the presence of these redundant records in the test set causes the evaluation results to be biased by the methods that have better detection rates on the frequent records [5]. New NSL-KDD datasets were generated in order to solve the issues of the original KDD dataset [5], with new train and test sets (KDDTrain+ and KDDTest) consisting of selected records of the complete KDD data set. This new version of the KDD data set (NSL-KDD) is publicly available for researchers [7]. According to Tavallaee et al. and McHugh, the major disadvantage of NSL-KDD is that it does not represent actual existing networks and associated attack scenarios [8, 5].

On the other hand, the combination of classifiers for IDS has been an effective research area for several years, and many research studies have concentrated on dealing with improving the accuracy of the proposed model. The most popular classifiers were the meta classifiers: Boosting, Bagging, and Stacking. Several classification techniques and machine learning algorithms have been tested by Choudhury and Bhowal to categorize network traffic, and out of several classifiers they concluded that Random Forest and BayesNet were suitable for intrusion detection, particularly when using the Boosting method [9]. A network anomaly detection strategy using an ensemble of three-based classifiers (C4.5, Random Forest, and CART) and Particle Swarm Optimization (PSO) for feature selection was proposed, which showed promising detection accuracy and a lower positive rate in contrast to existing ensemble techniques [10].

In 2015, Moustafa et al. generated a new dataset UNSW-15 in order to counter the unavailability of network benchmark data set challenges [11]. This data set contained a fusion of actual modern normal network traffic and contemporary synthesized attack activities thereof. The authors were able to find only one paper that described the use of different existing machine learning classifiers to evaluate complexities in terms of accuracy and false positive rate algorithms on the UNSW-15 dataset [12]. The results indicated that the Decision Tree classifier accomplished the