Anomaly Detection in Networks Using Machine Learning

Nagavinay kongara

1.Introduction

Motivation

Every day millions of people and hundreds of thousands of institutions communicate with each other over the Internet. In the past two decades, while the number of people using the Internet has increased very fast, today this number has exceeded 4 billion and this increase is continuing rapidly day.Parallel to these developments, the number of attacks made on the Internet is increasing day by Against these attacks, there are two basic methods used to detect the attacks in order to ensure information security; identification based on signature, and detection based on anomaly. Signature-based methods use the database they created to detect attacks. This method is quite successful, but the databases need to be constantly updated and new attack information processed. Moreover, even if the databases are up-to-date, they are vulnerable to the zero-day (previously unseen) attacks. Since these attacks are not in the database, they cannot prevent these attacks. The anomaly-based approach focuses on detecting unusual network behaviours by examining network flow. This method, which has been successful in detecting attacks that it has not encountered before, so is effective against zero-day attacks[1, 3]. In addition, more than half of today's internet usage is encrypted using SSL / TLS (Secure Sockets Layer / Transport Layer Security) protocols, and this rate is increasing day by day[4]. Because of the inability to observe the contents of the encrypted internet stream, signaturebased methods do not work effectively on this type of data. However, the anomaly-based approach analyses data by its general properties such as size, connection time, and number of packets. So, it does not need to see the message content and it can also do the analysis of encrypted protocols. Due to all these advantages, the anomaly-base detection method is being used intensively to detect and prevent network attacks.

Goals and Objectives Goals

- The goals that are aimed to achieve at the end of this study are as follows [1]: • Examination of machine learning algorithms that can be used to detect network anomalies.

• To detect network attacks in a fast and effective way by studying network anomaly with machine learning methods.

- • To determine the success level of the study by comparing the results obtained in it with the studies previously conducted in this area. • Contributing to the literature by obtaining close results from previous studies in the detection of network anomalies.

Objectives

The objectives that are aimed to achieve at the end of this study are as follows [1]: • To examine the previous work done in the field by doing extensive field research.
• Selecting the appropriate dataset by performing comprehensive research on the alternatives to the dataset.
• Choosing suitable algorithms by conducting extensive research on machine learning algorithms.
• Deciding on right algorithms by performing exhaustive research on machine learning methods.
• Selecting the appropriate software platform.
• Choosing the suitable hardware/equipment platform

• Deciding on the right evaluation criteria.
• Choosing the benchmark studies to be compared during the evaluation phase.

Structure of The Report

In the Background and Related Work section, some preliminary information is provided for a better understanding of the work such as data sets, anomaly and attack types, detailed information about attacks and machine learning algorithms. Finally, similar studies in the literature are listed and given information about them. The Methodology section consists of two subsections. The first subsection, Tools and Methods, provides information about the software and hardware used during the work and describes the performance evaluation methods. The second step, the implementation, contains these sub steps; data cleaning, the division of the data into training and testing, implementation of feature selection and machine learning algorithms. In the Results and Discussion section, the results obtained in the implementation step are shared and their cause-effect relationships and alternative approaches are discussed. In the Evaluation section, the data obtained in the study is compared with another study selected from the literature. 3 In the Conclusion and Future Work section, the study is briefly summarised and possible developments and innovation which could be applied in the study are touched.

Background and Related Work

Datasets

In the detection of network anomaly by machine learning methods, there is a need for a large amount of harmful and harmless network traffic for training and testing steps. However, it is not possible for real network traffic to be used publicly because of privacy issues. To meet this need, many datasets have been produced and continue to be produced. In this step, information will be given about some popular datasets, after that, they will be compared and evaluated to decide which of them to use in the implementation phase.

DARPA 98

With this dataset created by MIT Lincoln laboratory with DARPA funding, it is aimed to create a training and testing environment for Intrusion Detection Systems. In this dataset, the United States Air Force's local computer network is simulated. The data stream consists of processes such as file transfer via FTP, internet browsing, sending and receiving e-mail and IRC messages. In addition to Benign/Normal network traffic, it includes 38 attacks that can be grouped under attack types such as Denial of Service (DoS), User to Remote (U2R), Probe, and Remote to Local (R2L)_[6]. This dataset has received a lot of criticism, especially not including the fact that it does not reflect.

CAIDA [10]

(Centre of Applied Internet Data Analysis) is an organisation engaged in internet data analysis. The dataset provided by the facilities of this organisation is referred to by the same name. The dataset that makes up this dataset comes from a few hours of data flow recording of the OC48 backbone connection over San Jose city. This dataset also contains a section that simulates an hourly DDoS attack [7]. In the CAIDA dataset, data flows are exemplified only by specific applications and specific attacks. Therefore, the variety of sampling is quite limited. In addition, in this data set, data streams are not labelled. The fact that the data are unlabelled makes it very difficult to use this dataset in machine learning applications [11].

NSL-KDD [12]

Although the KDD99 dataset created in 1999 was a very good alternative to DARPA98, it was observed that there are too many repetitions in KDD99, and these repetitions affect the results of the studies performed with it and the performance of the machine learning algorithms. In addition, because the size of KDD99 is too large, researchers have tried to use part of this data set. Unfortunately, in order to minimise the dataset, randomly selected data from within could not capture all the properties of the data set [9]. To overcome these shortcomings, in 2009, Tavallaee et al. [13] created a new data set called NSL-KDD. In this version, they eliminated the mistakes and repetitions in KDD99. The NSLKDD dataset consists of 4 parts under two main headings as training and testing: training data (KDDTrain+), 20% of the training data (KDDTrain+_20Percent), test data (KDDTest+) and a smaller version of the test data with all difficulty levels(KDDTest-21)[12].

Anomaly and Attack types

2.1 Anomaly Types

The anomaly is a sample that does not have well-defined properties of a normal sample[11]. In order for the anomaly to be understood, the rules that make up the normal concept must be specific and valid. The anomaly is analysed under 3 headings:

Point anomaly: If a particular data sample looks different from the whole dataset with the properties it carries, it is called a point anomaly [11]. For example, it is a point anomaly that a person who makes a low amount of shopping every day with a credit card makes a very high amount of shopping on a random day.

Contextual anomaly: The out-of-pattern behaviour of a data sample depends on certain conditions or occurs under certain conditions [11]. For example, it is a contextual anomaly that a person who makes a low amount of shopping every day with a credit card makes a very high amount of shopping on the feast day.

Conclusion and Future Work Conclusion

In this study, it is aimed to detect network anomalies using machine learning methods. In this context, the CICIDS2017[16] has been used as dataset because of its up-to-dateness, wide attack diversity, and various network protocols (e.g. Mail services, SSH, FTP, HTTP, and HTTPS)[4]. This dataset contains more than 80 features that define the network flow. During the application, the importance weight calculation was made with the Random Forest Regressor[63] algorithm to decide which of these features will be used in machine learning methods. Two approaches have been used when making these calculations. In the first place, importance weights are calculated separately for each attack type. In the second method, all the attacks are collected under a single group and the importance weights for this group are calculated. that is, the common properties that are important for all attacks are determined. Finally, seven machine learning algorithms, which are widely used and have different qualities, have been applied to this data. These algorithms and the achieved performance ratios according to F-measure are as follows (F-measure takes a value between 0 and 1): Naive Bayes: 0.86, QDA: 0.86, Random Forest: 0.94, ID3: 0.95, AdaBoost: 0.94, MLP: 0.83, and K Nearest Neighbours: 0.97.

Future Work

This work is open to future improvements. In this section, a few of these improvement possibilities will be presented. In this study, a data set consisting of CSV files containing features obtained from the network flow was used as the training and test data. Unfortunately, this method is not practically viable in real systems. However, this problem can be solved by adding a module that catches real network data and makes it workable with the machine learning algorithm. Another point was that during this study, various machine learning methods were applied independently from each other and experimental results were obtained. However, this method has a weak practical applicability in real life. In order to deal with this problem, a multi-layered / hierarchical machine learning structure can be designed. Moreover, thanks to such a structure, it is possible to save time, CPU power and memory. For example, in a two-tiered structure, the 44 first layer can be created from fast and computationally cheap algorithms such as Naive Bayes or QDA, so network traffic can be observed continuously and at minimal cost. The first step, when detecting any anomaly, transmits it to an upper layer composed of algorithms with higher performance level. This layer measures by forming the decision mechanism. The first step, when detecting an anomaly, transmits it to an upper layer composed of algorithms with higher performance such as ID3, AdaBoost, and KNN. The final layer that makes up the determination mechanism takes the precautionary decision to protect the network against attack.