Network security prediction model using neural networks

Raghav Venkatesh and Raja Muthalagu

Department of Computer Science, Birla Institute of Technology and Science Pilani, Dubai Campus, Dubai, UAE.

Abstract. The aim of this paper is to design a robust network security prediction model. The designed model is an intrusion detecting model constructed using neural networks. The intrusion detecting model detects anomaly and misuse-based attacks. Intrusion detecting model also performs three kinds of classification tasks. The tasks include classifying between occurrence of an attack or a normal case, classifying between different attack types or a normal case and classifying between occurrence of an attack or other case. The intrusion detecting model also shows the classification accuracy, execution time and amount of memory usage. The objectives of the intrusion detecting model are high accuracy, low execution time and minimum amount of memory usage. The intrusion detecting model constructed using neural networks, fulfils the objectives of high accuracy, low execution time and minimum amount of memory usage.

1. Introduction

In today's world, networks are becoming more complex, interconnected and more widely used than ever before. The network traffic is almost exponentially increasing today and there is high demand for network usage as well. The networks are also becoming more vulnerable to being attacked by hackers or anyone with malicious intentions to destroy network systems. Vulnerable networks are in danger of crippling the economy and destroying privacy of all the people in this world. Thus, there is a need to improve the mechanisms to detect vulnerabilities of the network and improve network security prediction which is developed in the following model.

The following network security prediction model also has an aim to reduce memory consumption as well as improving detection of various types of attacks in terms of time and accuracy. The successive sections of this paper include previous work, intrusion detecting model using neural network section, results, conclusion and future work followed by references, at the end.

Algorithm and user roles section comprises of 4 different types of algorithm that were first hypothesised as probable tools to be utilized in designing the model. The algorithms mentioned in this section are support vector machine(SVM), Fuzzy Clustering, K-Means Clustering and Apriori. There is a brief overview of concepts involved in algorithms as well as various proposals over a period that helped in improvising and improving working of algorithms. This section also details 4 major network security user roles in a network. The 4 major network security user roles are data collectors, data providers, data miners and decision makers. Their objectives of preserving sensitive data are detailed.

Intrusion detecting model using neural network section comprises of four parts which are data stipulations, data pre-processing, usage of methods and implementation. Data stipulation sub-section details creation of files which are used for detection of anomaly-based assaults as well as misuse-
based assaults. There are also details of number of attack or normal instances for process of detection of anomaly or misuse-based assaults.

Data pre-processing sub-section details manner in which extra data that is unnecessary or of very less utility, is removed so that there can be optimum amount of data for classification. Thus, performance of network in detection process improves. Usage of methods sub-section consists of creation of corresponding datasets for anomaly and misuse detection. There is information about ‘neuralnet’ package which was used for building neural network. This neural network was used for detection of various attacks as well as classification.

Implementation sub-section comprises of 10 different attack types that were chosen. The 10 different attack types are Back, Buffer Overflow, FTP Write, Guess Password, Neptune, NMap, Normal, Port Sweep, Rootkit, Smurf and Satan. In this sub-section, there is mention of various classification processes that take place in anomaly detection attacks, misuse detection attacks and individual attack types. Also, there is mention of parameters of results such as classification accuracy, time taken for processing and resource consumption.

Results section consists of results of anomaly detection attack, misuse detection attack and individual attack types. Results include classification accuracy, time taken for processing and resource consumption. The values in results and the significance with regards to classification is also mentioned in this section.

In anomaly detection attack sub-section, there is classification of occurrence of an attack or a normal case. In misuse detection attack sub-section, there is classification between 10 different attack types and a normal case. In individual attack type sub-section, there is classification between a particular attack and other instances.

2. Algorithm and User Roles

There are four algorithms discussed in this section namely SVM algorithm, Fuzzy Clustering Algorithm, K-means Clustering Algorithm and Apriori algorithm which are explained below. This is followed by detailing of 4 different network user security roles and their access controlling objectives.

2.1. SVM
This algorithm is used to solve classification problems. Based on the basic construction of statistical principle, a kernel function is added to the calculation process to map low-dimensional problem to high-dimensional space, and finally the optimal solution in high-dimensional solution space is obtained. This means using SVM algorithm unlocks hidden patterns in a large amount of data, so as to discover the information behind the data. After uploading information to the system, it can identify time series or development trend of the data and make accurate judgments [3].

There was also a proposal regarding the use of MIES method to automatically gain the most optimal Gauss kenel parameters to obtain best hypo sphere [5]. There was an improved version wherein the algorithm of K-method changed into a combination with Apriori to achieve correct detection value of Root to Learn and User to Root at KDDCUP99 information set through 98% and 79% [6]. The idea of usage of dimensionality discount set of rules mixed with One-class SVM was proposed and received eleven fields of pre-processed records [7].

2.2. Fuzzy Clustering Algorithm
The fuzzy clustering analysis set of rules process is as follows [1]:
- Determining a similarity function
Establishing a corresponding fuzzy similarity matrix according to similarity function.
Calculating the fuzzy relationship and usage of flat method. Also includes forestalling when finding the transitive closure.
Classification in line with extraordinary thresholds and obtaining specific dynamic clustering effect.
Degrees are grouped together into a candidate assault series set.
Finally, collection sample mining algorithm is used to mine assault mode of the hacker from candidate assault sequence.

A new fuzzy rule generation approach became a proposal in which clusters in the training pattern set in keeping with the fuzzy C-method clustering technique, in line with characteristics of each pattern and the cluster [8].

2.3. K-means Clustering Algorithm
The K-means clustering algorithm assumes that the required clustering ok values are known, however in fact inside the security analysis, the k values are typically unknown. And the selection of the initial clustering centre of K-way set of rules is important. Mini Batch K-Means is used to divide normal dataset and attack dataset into clusters with similar size separately and centre of each cluster is used as cluster index. Then there is a method to select representative instances from each cluster. The representativeness of an instance is related to both density and distance. Higher representativeness makes an instance more likely to be chosen as a representative instance. After selection, there is assigning of a weight to each representative instance. This step not only reduces the size of the original data, but also preserves maximum amount of information [2].

All the models have a common aim of detecting vulnerabilities in a network system more accurately more efficiently and more quickly. Different algorithms have been proposed to achieve the above-mentioned goal [2]. As early as 2009, an intrusion detection method based totally on K-manner algorithm was proposed [9]. A hybrid intrusion detection algorithm based on K-method and selection tree was proposed [10]. A data function screening technique became used to pre-procedure the 41-dimensional functions in the statistics set [11].

2.4. Apriori Algorithm
Apriori algorithm is used for internal rule association mining of security facts because it has high-quality significance. Frequent scanning of transaction database and immoderate set of sub-waiting options is a problem of this algorithm [12,13]. The minimal support and minimum confidence values have a massive impact on detection outcomes. The Apriori algorithm was proposed for the first time ever in 1993 [12].

2.5. Different User Roles
There is a necessity to ensure that sensitive data does not get leaked to those with malicious intentions. Due to this, there are specific access controlling objectives for different network security user roles. There are 4 such major user roles. They are data providers, data collectors, data miners and decision makers [4]. Also, it is crucial to make sure that sensitive facts do not get leaked to those with malicious intentions. Due to this, there are access controlling objectives for extraordinary network safety consumer roles. There are 4 such primary user roles. They are information providers, information collectors, information miners and decision makers [4]. For data providers, access controlling objective is to efficaciously control the quantity of sensitive data revealed to others. To achieve this goal, there can be utilization of protection tools to restrict other’s get right of entry to their information, promote the data at auction to get enough compensations for privacy loss, or falsify information to hide their genuine identity.
For data collectors, access controlling goal is to launch useful facts to facts miners without disclosing records providers’ identities and sensitive statistics approximately them. To achieve this goal, there is a need to develop right privacy models to quantify possible loss of access control underneath exceptional attacks, and practice anonymization strategies to the statistics [4]. For data miners, privacy-maintaining objective is to get correct records mining outcomes while keeping the sensitive statistics undisclosed either within the manner of records mining or within the mining outcomes. To attain this goal, proper approach may be chosen to regulate the information before positive mining algorithms are carried out. Also, steady computation protocols may be utilized to ensure safety of private data and sensitive statistics contained within learned model. For decision makers, access controlling objective is to make a correct judgement, approximating credibility of the facts mining results they get. To attain this goal, provenance techniques can be utilized to hint the returned records of received facts or build classifier [4].

3. PROPOSED SYSTEM

3.1. Data Stipulations
Two documents are created; one to detect anomaly-based assaults and another to detect misuse-based assaults. These documents had approximately 4500 records. The inputs are divided into a Training Data Set (75%) to train Neural Network and Test Data Set (25%) on trained Neural Network.

3.2. Data Pre-processing
The first goal in method was simplifying facts to be processed. Simplifying implies doing away with attributes that are of less utility. The benefit is getting rid of attributes reduces the dimensions of data getting processed, which improves performance of the neural network. The drawback is if crucial attributes are accidently removed then accuracy of detecting an intrusion will suffer. All constant attributes had been removed and attributes that passed the most percent of a variance parameter were additionally removed. “R” scripts also checked satisfactory of a characteristic primarily based on variance inside characteristic samples. Attribute discount was carried out for attributes that do not contribute even 1 percent of cumulative variation in the facts set.

3.3. Usage of Methods
For wearing out intrusion detection for Anomaly primarily based attacks and Misuse primarily based attacks there have been two files dataset of anomaly and dataset of misuse. In anomaly detection information set, the class or prediction variable was either normal which represented an everyday case or an attack. Misuse detection facts set had a category variable Normal or Name of the assault which represents a specific kind of assault such as Smurf, NMap, Rootkit, etc.

The data cleaning was achieved on files consisting of dataset anomaly and dataset misuse. Usage of Weka to gain dataset anomaly’s attribute selection and dataset misuse’s attribute selection meant much lesser attributes that assisted in speeding up the NN. The ‘neuralnet’ package is available in R and is open source. It became used for neural network-based IDS and analysis. The package deal provided capabilities to both generate neural network and carry out classification.

3.4. Implementation
For this exercise, there was a consideration over 4500 instances of normal cases and attack instances. 10 types of attacks including Neptune, NMap, PortSweep, Satan, Smurf, BufferOverflow, FTPWrite, GuessPassword, Back and Rootkit attacks were chosen. In the procedure of anomaly detection, there has been script implementation of anomaly primarily based Intrusion Detection to offer confusion matrix, class accuracy, time taken for implementation and resource consumption. There was classification if there was any attack or normal case. For detection of misuse attacks, script implemented primarily misuse based intrusion detection to offer confusion matrix, classification accuracy, time taken for implementation and resource consumption. There was a classification among 10 attacks and normal case. For categories of individual attacks, the script carried out misuse detection
system capable of detecting an attack to provide confusion matrix, class accuracy, time taken for implementation and resource consumption for 10 attack types. There was classification among a selected attack and other instances.

4. Results and Discussions

4.1. Anomaly detection attack
In the set of results shown in table 1, there are details of accuracy while detecting anomaly attack, execution time and the amount of memory consumption. [14]. There is classification of occurrence of an attack or normal case in this process. Values mentioned in table 1 indicate number of records. The value of (Attack, Attack) coordinates and value of (Attack, Normal) coordinates are 389 and 6 respectively. Value of (Attack, Attack) coordinate is higher than (Attack, Normal) coordinate as shown in Table 1. This implies occurrence of an attack. The classification accuracy is high at 99.57 percent. Execution time is minimal at 3.9979 seconds. Memory usage is minimal at 2191.311 Kbs as shown in Table 2.

| Table 1. Anomaly detection attack – No. of Record details |
|------------------------------------------------------------|
| Axis 1 | Anomaly attack | Normal |
|--------|----------------|--------|
| Anomaly attack | 389 | 6 |
| Normal | 3 | 763 |

| Table 2. Anomaly detection attack results |
|-------------------------------------------|
| NSPM Accuracy: 99.57% |
| Execution Time: 3.9979 seconds |
| Memory Usage: 2189.311 Kbs |

4.2. Misuse detection attack
In the set of results shown in table 3, there are details of accuracy while detecting misuse attack, execution time and the amount of memory consumption. [14]There is classification between 10 attack types and normal cases. In table 3, just as anomaly detection case, there are 2 axes, i.e. axis 1 and axis 2. Values mentioned in table 3 indicate number of records. The values of (Back, Back), (Buffer Overflow, Buffer Overflow), (Guess Password, Guess Password), (Neptune, Neptune), (Nap, NMap), (Port Sweep, Port Sweep), (Satan, Satan) and (Smurf, Smurf) coordinates are highest among all row values (values are 67, 4, 12, 57, 75, 748, 63, 59 and 58 respectively). This indicates easier classification and greater possibility of an attack. (FTP Write, FTP Write), (Rootkit, Rootkit) coordinates are significantly low (values are 1 and 0 respectively). Values of (FTP Write, Normal) and (Rootkit, Normal) coordinates are 0 and 1 respectively, which are not the highest row values. Thus, there is no occurrence of normal case as shown in table 3.

As shown in table 4, the classification accuracy is high at 98.1%. Execution time of 48.9282 seconds is higher than previous scenario due to larger amount of information, however it is still low for information processed. Memory usage of 2988.14 Kbs, is increased due to larger information, but not extremely high.
### Table 3. Misuse detection attack – No. of Record details

| Axis1      | Back | Buffer Overflow | FTP Write | Guess Password | Neptune | NMap | Normal | Port Sweep | Rootkit | Satan | Smurf |
|------------|------|-----------------|-----------|----------------|---------|------|--------|-----------|---------|-------|-------|
| Back       | 67   | 0               | 0         | 0              | 0       | 0    | 0      | 0         | 0       | 0     | 0     |
| Buffer Overflow | 0   | 4               | 0         | 0              | 1       | 1    | 0      | 0         | 0       | 0     | 0     |
| FTP Write  | 0    | 0               | 1         | 1              | 0       | 0    | 0      | 0         | 0       | 0     | 0     |
| Guess Password | 1  | 0               | 12        | 0              | 1       | 0    | 0      | 0         | 0       | 0     | 0     |
| Neptune    | 0    | 0               | 0         | 0              | 57      | 0    | 0      | 1         | 0       | 0     | 3     |
| NMap       | 0    | 0               | 0         | 0              | 0       | 75   | 0      | 0         | 0       | 0     | 0     |
| Normal     | 0    | 0               | 0         | 0              | 0       | 0    | 748    | 0         | 1       | 0     | 0     |
| Port Sweep | 0    | 0               | 0         | 0              | 0       | 0    | 63     | 0         | 0       | 1     | 1     |
| Rootkit    | 0    | 0               | 1         | 0              | 3       | 0    | 1      | 3         | 0       | 0     | 1     |
| Satan      | 0    | 0               | 0         | 0              | 3       | 0    | 1      | 1         | 59      | 1     | 1     |
| Smurf      | 0    | 0               | 0         | 0              | 0       | 0    | 0      | 0         | 0       | 0     | 58    |

### Table 4. Misuse detection attack Results

|             | NSPM Accuracy: 98.10% |
|-------------|-----------------------|
| Execution Time: | 48.9282 seconds     |
| Memory Usage:     | 2988.14 Kbs          |

The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections the formatting shown in table 2 should be used.

### 5. Conclusion and Future Work

In the network security prediction model, memory consumption was low, time taken for detection of the attacks is also low. Also, the accuracy of the detection of attacks is high. The above methods used to design the model is also simpler to design. The above methods are also far more cost effective as usage of neural networks in R is freeware. Also, computations are easier through usage of this model. Hence using neural network page in R is also an effective way for designing a network security prediction model. Therefore, the usage of neural networks using R is recommended for designing any type of a network security prediction model. The future scope is in designing models possessing capability of detecting any intrusions even more accurately and quickly with more minimal memory consumption during processing.
REFERENCES

[1] Zichuan J Yanpeng C Zheng Y 2019 Survey of intrusion detection methods based on data mining algorithms *Proc. Int. Conf. on Big Data Engineering (New York)* pp 98-106

[2] Qihua W Xiaoqin O Jiacheng Z 2019 A Classification algorithm based on data clustering and data reduction for intrusion detection system over big data *KSII Trans. on internet and information systems* **13** pp 3714-3732

[3] Xiaoyi H 2020 Network security situation prediction based on grey relational analysis and support vector machine Algorithm *Int. J. of network security* **22** pp 177-182

[4] Lei X Chunxiao J Jian W Jian Y, Yong R 2014 Information security in big data: privacy and data mining *IEEE Access* **2** pp 1149-1176

[5] Xiao Y Wang H Xu W 2015 Parameter selection of gaussian kernel for one-class SVM *IEEE Trans. on Cybernetics* **45** pp 927-939

[6] Song C Ma K 2009 Design of intrusion detection system based on data mining algorithm *Int. Conf. on signal processing systems* (Singapore) pp 370-373

[7] Rochim A F Aziz M A Fauzi A 2019 Design log management system of computer network devices infrastructures based on ELK stack *Int. C. on Electrical Engineering and Computer Science (ICECOS)* (Batam Island) pp 338-342

[8] Jin C, Ye Z, Wang C, Yan L Wang R 2018 A network intrusion detection method based on hybrid rice optimization algorithm improved fuzzy C-means *IEEE 4th Int. Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems* (IDAACS-SWS) (Lviv) pp 47-52

[9] Jianliang M Haikun S Ling B 2009 The application on intrusion detection based on K-means cluster algorithm *Int. Forum on Information Technology and Applications* (Chengdu) pp. 150-152

[10] Aung Yi Y Myat M M 2018 Hybrid intrusion detection system using K-Means and classification and regression trees algorithms *IEEE 16th Int. C. on Software Engineering Research, Management and Applications* (SERA) (Kunming) pp 195-199

[11] Ravale U Marathe N Padiya P 2015 Feature selection based hybrid anomaly intrusion detection system using K-Means and RBF kernel function *Procedia Computer Science* pp 428-435

[12] Han J W Kamber M P 2011 Data mining: concepts and techniques (3rd Edition. Elsevier Science)

[13] Huang Z 2019 Research and implementation of intrusion detection based on host log”. *Proc. of the Int. Conf. on Big Data Engineering*, pp 98–106

[14] Pratik M, Saunil D, Ravina D 2015 Intelligent Security Systems- Intrusion Detection System https://github.com/jgera/Network-Intrusion-Detection-System/blob/master/report.pdf