A Data Analytics System for Network Intrusion Detection Using Decision Tree

Terungwa Simon Yange1,*, Oluoha Onyekwere2, Yakubu Musa Abdulmuminu3

1Department of Mathematics/Statistics/Computer Science, University of Agriculture, Makurdi, Nigeria
2Department of Computer Science, University of Nigeria, Nsukka
3Department of Computer Science, Federal University Lokoja, Lokoja, Nigeria
*Corresponding author: lordesty2k7@ymail.com

Received April 26, 2020; Revised May 28, 2020; Accepted June 04, 2020

Abstract Network intrusion detection systems are becoming an important tool for information security and technology world. Given the rise of attacks across the network, there is a pressing need to develop an improved security system to combat these growing threats on the computer network. The quality of an intrusion detection system is determined by the number of attacks its able to classify correctly. This research developed a data analytics system for network intrusion detection to combat the ever growing threats as well as classify them so as to ease the task of data scientists and network administrators. Decision tree algorithm and python programming language were used. KDD’99 was used as the data source. Decision tree assists the network administrator to decide about the incoming traffic, i.e., whether the coming data is malicious or not by providing a model that separates malicious and non-malicious traffic. It allows taking less number of attributes and provides acceptable accuracy in reasonable account of time. From the results of the experiments, it is concluded that the system is more efficient with respect to finding attacks in the network with less number of features and it takes less time to construct the model. Also, the efficiency of the system has little or no regards for the size of the dataset and the number of features used to construct the decision tree.

Keywords: data, analytics, decision tree, intrusion detection, attack, intruder


1. Introduction

Over the years, network intrusion detection systems have attracted the attention of many researchers. This is due to the fact that computer networks are widely applied in many spheres of life endeavours which include business, medicine, engineering and other fields. This has necessitated the building reliable networks whose absence is inimical to human existence. Although the rapid advancement in information technology have created large and complex network which has given birth to myriad of challenges leading to difficulties in building reliable networks. Network intrusion detection system seems to be the feasible solution against malicious attacks on networks as it boosts security of these networks. Due to the dynamic and ever changing nature of attacks, several experiments have been conducted and using different techniques [1].

Intrusion is defined as any set of actions that attempt to compromise the integrity, confidentiality or availability of system resources. Any real world entity that tries to find a means to gain unauthorized access to information, causes harm or engage in other malicious activities is known as intruder or an attacker. The intrusion of computer network has been on a high side and this is cause by unauthorized users of systems who attempt to gain further privileges for which they are not authorized, or authorized users who are guilty of user policy violation [2]. Intrusion detection therefore involves the gathering of intrusion associated information while monitoring and analysing the events triggering these attacks in a computer network so as to ascertain the intruders. Most intrusion occurs via network utilising network related protocols to attack their targets. Though there are many types of attacks threatening the availability, integrity and confidentiality of computer networks but the denial of service attack (DoS) is considered as one of the most common harmful attacks [3]. Data analytics on the other hand is an important concept in the world of computing that involves the process of modelling data with the aim of actualizing structured information and supporting decision making system.

The application of data analytics in the detection of intrusion in computer networks therefore involve the process of monitoring, protecting, capturing and analysing data to detect malicious activities in computer networks. This work by collecting information from different sources within the computer systems and network, compares them to pre-existing patterns to verify if they are attacks or malicious activities. A network intrusion
detection system relates an intrusion to the system once it has taken place and signals an alarm. It primarily watches out for attacks that occurs or starts from within the system. It allows organization to protect their systems from threats that come with increasing network connectivity and reliance on information systems. 

With the rapid growth in the sizes of computer networks and adoption of web applications, there has been significant increase in the potential damage that comprised the security of networks. This damage includes web site defacement, corruption and loss of data, denial of services, viruses, Trojans and worms. In fact, network security involves three realities: first, the defender has to defend against every possible attack, while the attacker only needs to find out one weakness; secondly, the immense complexity of modern networks makes it impossible to be properly secure; and finally, professional attackers may encapsulate their attacks in programs, allowing ordinary people to use them. Vulnerability assessments and intrusion prevention or intrusion detection are just one aspect of IT security management. However, due to recent developments with the continuing spread of network connectivity IT security management, is faced with yet another challenge, requiring a structured approach for an adequate response. Many existing intrusion detection systems do not consider attacks within the network [2]. These systems only consider incoming attacks otherwise known as external attacks with patterns in attack databases or check for strict deviation from the normal network traffic. The systems analyse and monitor attacks on the network by capturing the data or packets entering the network, compares it to the attacks in an existing database or compares data or packets entering the network, if there is no match or an abnormality then it is considered as an attack on the network. The major problem associated with these existing systems is that an attack coming from within the computer or network is hardly detected and this has negatively exposed them to many internal attackers.

It is on the basis that this research is carried out. It built a system that is capable of analysing and protecting data by classifying known and unknown attacks or malicious behaviours within a computer system in a network. It utilised the decision tree algorithm to analyse intrusion within a network using the KDD99 dataset. The system monitors the packets’ behaviour from time to time and report when some intrusions are detected. This is very important in network defence processes and aids system administrators in providing information about malicious attacks within a network.

2. Literature Review

As Internet services spread globally, security threats have also continue increased and this has also rise the demand for intrusion detection systems. Intrusion detection systems are vital in network defence processes and help system administrators to be forewarned about incoming attacks [1]. According to [4] network intrusion detection system is categorized into: Misused Detection and Anomaly Detection.

**Misused Detection:** This suggests patterns of existing attacks to detect intrusions. Misuse detection is based on the knowledge inferred from the pattern of previous attack. It attempts to recognize the intrusion pattern that has been recognized and stated earlier. It cannot detect unknown errors and function well with offline data.

**Anomaly Detection:** This check for strict deviations or noise from the normal network traffic and report it as attack or a malware. Thus, it can detect both known and unknown attacks. Anomaly detection systems work better with online data.

2.1. Application Based IDS

Application Based IDS (APIDS) will check the functional behaviour and event of the protocol. The system or agent is placed between a process and group of servers that monitors and analyses the application protocol between devices. Intentional attacks are the hostile attacks carried out by malcontent employees to cause harm to the organization and Unintentional attacks causes financial damage to the organization by deleting the important data file. There are numerous attacks have been taken place in OSI layer [5].

**Denial-of-Service (DOS) Attacks:** DOS refers to Denial-of-Service and is best defined as an attempt to make a computer(s) or network(s) unavailable to its intended users or also a Denial of Service attack is when an attacker is trying to generate more traffic than you have resources to handle.

**DOS and DDoS:** In a DOS attack, one computer and one internet connection also is established to overwhelm a server or network with data packets, with the only intention of overloading the bandwidth of victim and available resources. A Distributed Denial of Service (DDOS) attack is the same, but it is amplified. Rather than one computer and one internet connection a DDOS is, and often involves millions of computers all being used in a distributed manner to have the effect of hitting a web site, web application or network offline.

In both cases, either by the DOS or the DDoS attack, the target is bombarded with data requests that have the effect of disabling the functionality of the victim.

**SYN Attack:** SYN attack is also defined as Synchronization attack. Here, the attacker sends the flood of SYN request to the destination to use the resources of the server and to make the system unresponsive.

**Peer-to-peer attacks:** A peer-to-peer or P2P network is a distributed network in which individual nodes in the network called “peers” act as both suppliers (seeds) and consumers (leeches) of resources, in contrast to the centralized client– server model where the client server or operating system nodes request access to resources provided by central servers.

**Ping of Death:** A type of DOS attack in which the attacker sends a ping request that is larger than 65,536 bytes, which is the maximum size that IP allows onto the network. While a ping larger than 65,536 bytes is too large to fit in one packet that can be transmitted through, TCP/IP allows a packet to be fragmented, essentially splitting them in smaller segments that are reassembled at the end. Attacks took advantage of this limitation by
fragmenting packets that when received packet would
total more than the allowed number of bytes and would
effectively cause a buffer overload on the operating
system at the receiving end then the system could crash.

**Eavesdropping Attack:** It is the scheme of interference
in communication by the attacker. This attack can be done
over by telephone lines, instant message or through email.

**Identity Spoofing (IP Address Spoofing):** Most
operating systems and networks use the IP address of a
computer to identify a valid entity on the network. In
certain cases, it is possible for an IP address to be falsely
assumed have spoofing identity. An attacker might also
use special programs to construct IP packets that are
originate from valid IP addresses inside the corporate
intranet. After gaining access to the network with a valid
IP address, the attacker can be modifying, rerouting, or
deleting your data.

**Man-in-the-Middle Attack:** As the name suggests, a
man-in-the-middle attack occurs when someone between
you and the person with whom you are communicating is
actively monitoring, capturing, and controlling your
communication transparently. For example, the attacker
can re-route a data exchange. When computers are
communicating at lowest levels of the network layer such
as physical layer, the computers might not have been able
to decide with whom they are exchanging the data. Man-
in-the-middle attacks are like someone assuming your
identity in order to read your message. The person on the
other end might believe as it is you because the attacker
might be actively replying as you to keep exchanging the
information.

**Application Layer Attack:** An application-layer attack
targets the application servers by intentionally causing a
fault in a server’s OS or applications. This results in the
attacker gaining the ability to bypass accessing normal
controls. The attacker takes advantages of this situation,
gaining control of your application, system, or network, and
can do any of the following:
- Read, add, delete, or modify your data or operating
  system.
- Can introduce a virus program that uses your
  computers and software applications to copy viruses
  throughout entire network.
- Can introduce a sniffer program to analyse your
  network and gain information that can be used to
  crash or to corrupt your systems and network.
- Abnormally terminate your data applications or
  operating systems and Disable other security
  controls to enable future attacks.

**Sniffer Attack:** A sniffer is an application or device
that can monitor, read, and capture network data exchanges
and read network packets. If the packets are not encrypted,
a sniffer provides a full view of the data inside the packet.

### 2.2. Network Intrusion Detection System

Using Neural Network

Artificial Neural Networks (ANN) are nonlinear
information processing devices, built from interconnected
elementary processing devices called neurons inspired by
the way biological nervous systems. The development of
the ANN started in 1943 by McCulloch and Pitts and is
still growing extravagantly. The advantages of ANN
include adaptive learning, self-organization, parallelism,
fault tolerance etc., applications involve in knowledge
extraction, pattern recognition, forecasting, clinical
diagnosis, security systems and still wider. Intrusion
Detection Systems based on wavelet and Artificial Neural
Network (ANN) that was applied to a popular Knowledge
Discovery in Data Mining (KDD) was presented by [6].
The results showed that approach suggested high detection
rate of malicious attacks.

Lidong [7] introduced Intrusion detection as an
important issue that needs a lot of attention in the
computer world and uses TDDNN neural network to
recognize the temporal behaviour of network attacks. The
system captures packets in real time using a packet
capture engine that presents the packets to a pre-
processing stage using two pipes. The pre-processing
stage extracts the relevant features for port scan and host
sweep attacks, stores the features in a tagged line of a
TDNN, and produces outputs that represent possible
attack behaviours in a pre-specified number of packets.
These outputs are used by the pattern recognition neural
networks to recognize the attacks, which are classified, by
the classifier network to generate attack alerts.

Ghosh et al. [8] in an attempt to detect internal
intrusion offered ANN-based intrusion detection on the
road to recognize and foretell unusual activities in the
system. The data for training and testing was gathered
from DARPA and the proposed neural network technique
for detecting intrusion produced significant results.

Manzoor and Kumar [9] approaches the challenge of
detecting attacks using network intrusion detection in a
two-fold manner. First, a fully connected Deep Neural
Network (DNN) was used to train a system with
supervised learning using labelled benign and malicious
network traffic data. Newer benchmark datasets produced
by the Canadian Institute of Cyber Security from the
University of New Brunswick (CIC UNB) were used
which are more representative of modern day network
traffic and attacks and do not have drawbacks of previous
datasets commonly used in the field. After learning these
patterns of malicious and benign by training a fully
connected neural network, the system can reliably and
effectively detect and classify modern attack traffic with a
high degree of accuracy, high rate of recall, and a low rate
of false positive rate. This is considered to be a form of
pattern-based detection.

### 2.3. Network Intrusion Detection Based on
Clustering

Clustering is the process of grouping a set of physical
or abstract objects in such a way that objects in the same
group are more similar to each other than to those in other
groups [10]. K-Means, canopy, DBSCAN, EM, Fuzzy,
C-Means, CLOPE, and cobweb are popular implementations
of clustering. The clustering techniques can be
implemented in several fields, such as image analysis,
information retrieval, bioinformatics, crime analysis, and
climatology. Azeem et al. [11] carried out an experiment
on network intrusion detection using KNN and Rough
Set. The both algorithms were compared based on
performance and concluded that the two algorithms
perform poorly if their representations are few in the
training dataset. However, the attribute values in a training data set completely differ from the attribute values from the test dataset mostly for these two attack types. This led to wrong classification because these instances were not learned in the training phase.

Rung-Ching et al. [12] proposed the implementation of an intrusion detection system using an SVM based system on a Rough Set Theory to reduce the number of features from 41 to 29. They also compared the performance of RST with Principal Component Analysis. The framework RST-SVM method result has a higher accuracy as compared to either full feature or entropy. The experiment demonstrated that RST-SVM yields a better accuracy.

Chibuzor and Bennett [13] developed a fuzzy class association-rule mining Genetic Network Programming (GNP) based method to detect network intrusions. GNP an evolutionary optimization technique, used directed graph structures instead genetic programming trees leading to enhanced representation ability with compact programs derived from reusability of nodes in a graph structure. By combining fuzzy set theory with GNP, the new method dealt with a mixed database of discrete and continuous attributes and extracted many important class-association rules contributing to improved detection ability.

Sharafaldin et al. [14] proposed a data mining technique to improve the detection rate of intrusion detection system. This clustering technique provide a better performance in Intrusion Detection accuracy rate, faster running time and detecting the false positive rate. To fragment a complex problem into sub problems for which the solutions obtained are simpler to realize, execute, supervise and update.

Verma et al. [15], carried out analysis using distance-based machine learning models to statistically analyse the complexity of CIDD9901 dataset used for evaluating anomaly based network intrusion detection systems. This study used two machine learning techniques for analysis i.e. kNN classification and k-means. For each pre-processed dataset file 66% of pre-processed data was used for building the model and 34% for testing.

2.4. Intrusion Detection Systems Using Decision Tree

Decision tree is used for predictive analysis where prediction is achieved by constructing a decision tree with test points and branches to classify or predict a couple of subgroup from interested object group by modelling rule and observing relation. At test points, decision is taken to pick classifier to pick a specific branch and traverse down the tree to reach to a final point of decision. This method models decisions and special forms of tree structure. Therefore, it has advantage to understand analysis process and results easily. Thuzar [16] proposed feature selection based on mutual correlation method to reduce the KDD99 dataset. The method was able to able to successfully differentiate between the four classes of attacks. Tiwari et al. [17] conducted experiments with Support Vector Machine and compared it with the Decision Tree algorithm performance using the KDD99 dataset. The experiment showed that decision tree is capable of multiclass classification which is not so with SVM. From the result, it was seen that Decision Tree gives better performance with small training data. The testing time and training time of the Decision Tree Classifier is better than that of SVM.

Nasimuzzaman et al. [18] performed several experiments to evaluate the efficiency and the performance of the following machine learning classifiers: J48, Random Forest, Random Tree, Decision Table, MLP, Naive Bayes, and Bayes Network. All the tests were based on the KDD intrusion detection dataset. The rate of the different type of the attacks in the KDD dataset was measured. Also, [19] carried out several experiments to evaluate the efficiency and the performance of some machine learning classifiers: J48, Random Forest, Random Tree, Decision Table, MLP, Naive Bayes, and Bayes Network. All the tests were based on the KDD intrusion detection dataset. The experiments demonstrated that there is no single machine learning algorithm which can handle efficiently all the types of attacks.

Kabir et al. [20] proposed an intrusion detection system that allows combining several decision trees which forms a classifier by measuring majority of votes of classifications specified by individual trees. This showed that the technique did well on live dataset. Yogendra and Upendra [21] reduced the features of the dataset using information gain of the attributes. They analysed four algorithms towards their suitability for detecting intrusions from the KDD99 dataset. The Decision tree classifier was found to perform more in terms of detecting intrusions and accuracy than the other classifiers.

2.5. Summary of Research Findings

This review revealed different network intrusion detection systems and to a greater extent, this have reduced the security challenges due to external attacks been faced in the information security world. Little or no research have been done on internal attacks which is the focus of this research. Minding the ever increasing security threat faced in the networks, it is important to build a system to help detect, classify and monitor systems against these attacks across and within the network. To this end, decision tree algorithm will be used because it works well with large data and it generalization accuracy to develop a Data Analytics Model for Network Intrusion Detection that would help network administrators.

3. Methodology

The methods deployed in achieving the aim of this research are captured in this section.

3.1. Data Collection

Data was collected from the KDD’99 datasets. KDD’99 datasets. It is considered as a standard benchmark dataset which include a wide variety of intrusions. The quality of the dataset was examined and evaluated using data mining and machine learning algorithms amongst others. A preview of the dataset used is shown in Figure 1. The dataset has 148517 records with contains network attack features such as HTTP, Duration, Flag, Service, SRC Bytes, Destination Bytes etc.
3.2. Architectural Design for the System

The architectural diagram is shown in Figure 2.

**Internet**: This provides access to other networks which attackers use to intrude other networks. This provides the data for external attacks.

**Data Pre-Processing**: This module collects data from the Internet and user modules. The data is recorded into a file and then pre-processed. This is where several computational methods are applied to organize and extract elements of the data packets, such as data cleaning, and data selection. It is a key step in the process, and it removes noise, outliers, and redundant or irrelevant information, handles missing data fields, and determines DBMS issues, such as types of data, schema, and the managing of missing and unknown values. The features are scaled to avoid features that may weigh on the result. Also, all categorical features are converted into numerical data. This is done using one-Hot-encoding. Then, Recursive Feature Elimination (RFE) is used to select features that are necessary for the model development. This is done using Univariate selection process. This analyzes each feature individually to determine the strength of the relationship between the feature and labels. When this subset is found, Recursive Feature Elimination (RFE) is applied. For instance, the Internet Protocol (IP) address of the source and destination system, protocol type, header length and size could be taken as a key for intrusion selection.

**Decision Engine**: The pre-processing phase is immediately followed by the analysis phase, where the decision tree algorithm and other related statistical inference methods are applied to make some deductions from the data. This phase helps to discover the most valuable feature of the data that is depending on the task and applies dimensionality reduction, such as reducing the number of attributes, attribute values, and tuples, or transformation methods, such as normalization, aggregation, generalization, and attribute construction, to reduce the effective number of variables under consideration or to find invariant representations for the data. Rule based IDS analyze the data where the incoming traffic is checked against predefined signature or pattern. Another method is anomaly based IDS where the system behavior is studied and mathematical models are employed to it. This is followed by the outcome, which uses probabilistic measures of confidence in the results for decision making. It defines about the reaction and attack of the system. It can either inform the system administrator with all the required data through an email/alarm icon or it can play an active part in the system by dropping packets so that it does not enter the system or close the ports.

![Figure 1. Sample of KDD'99 Dataset](image1)

<table>
<thead>
<tr>
<th></th>
<th>duration</th>
<th>protocol_type</th>
<th>service</th>
<th>flag</th>
<th>src_bytes</th>
<th>dst_bytes</th>
<th>land</th>
<th>wrong_fragment</th>
<th>urget</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>tcp</td>
<td>http</td>
<td>SF</td>
<td>181</td>
<td>5460</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>tcp</td>
<td>http</td>
<td>SF</td>
<td>239</td>
<td>486</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>tcp</td>
<td>http</td>
<td>SF</td>
<td>235</td>
<td>1337</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>tcp</td>
<td>http</td>
<td>SF</td>
<td>219</td>
<td>1337</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>tcp</td>
<td>http</td>
<td>SF</td>
<td>217</td>
<td>2032</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

![Figure 2. Architectural Representation of the System](image2)
Database: The database stores the processed data. At some points, the database is consulted before certain inferences are arrived at by the decision engine.

Users: This comprised of the users of the network. Packets from within are also monitored. They pre-processed and analysed by the pre-processor and the decision engine.

3.3. Application Algorithm

This is the step that describes the attributes of the various modules of the model. It shows the count, mean, minimum and other values of the attributes contained in the dataset. Decision Tree algorithm is based on C4.5 decision tree algorithm. The main issue in constructing decision tree is the split value of a node. The steps of the algorithm are as follows:

1. Collect the dataset, N
2. IF (A belong to class B)
   {Classify attack = B;
    Mark N as class C;
    Give decision;
    Return N;
   }
3. For I = 1 to n
4. Split dataset
5. Yb= training attributes
6. ya = testing attribute
7. N.ya = attributes having highest features
8. If (N.ya == continuous)
   {Find threshold}
9. For( Each A in splitting of A)
10. If(T is empty)
    {Attack feature is void ;}
    Else
    {child of N = is an attack category}
11. Build Decision tree classifier
12. Evaluate the classification error rate of Node N
13. Return N;
14. End

From the pseudocode above, to select the split value, C4.5 algorithm first sorts all the values of an attribute. Then from these sorted values, say, Ai, Ai+1, … An, the gain ratio of all the values is calculated by choosing the lower value of Ai and Ai+1 as threshold value and then calculate split value by using above mentioned formula. The value which gives the highest gain ratio is chosen as the split value for that particular node. Instead of using all these calculations which makes technique more complex and difficult to understand, we use a simple and effective approach. In our approach, there is no need to sort the attribute values to calculate the split value. We calculate the split value by taking the average of the values in the domain of a particular attribute at each node. It gives uniform weightage to all the values in the domain, making the classifier totally unbiased towards the most frequent values in the domain of an attribute. Sometimes, gain ratio may choose an attribute as a split attribute just because its intrinsic information is very low. This limitation can be overcome by considering only those attributes that have greater value of information gain than average information gain.

4. Results

The system was implemented using python programming language and the evaluation was carried out using confusion matrix to determine its accuracy and efficiency, and also to know how well and fast the system could use different data input. The experiments were performed on full training data set having 125973 records and test data set having 22544 records. First, we compute information gain of all the attributes of the data set. We found that there are 16 attributes whose information gain is greater than the average information gain. That’s why in the pre-process step, we can choose 16 or less than 16 attributes for further processing based on information gain because the remaining features will not have much effect on classification of the dataset. Then, the data set with these selected attributes is passed to the algorithm for constructing, training and testing the decision tree. Different attacks were considered two instances i.e., internal and external. The performance evaluation of denial of service attack of this model is shown in Figure 3. The Figure 4 shows the confusion matrix of the DoS and Probe attack categories. It’s a performance metrics used to test the validation of the system. The Figure 5 shows the accuracy of the U2R attack on the network. It uses the test data to check for the accuracy using the decision tree classifier.

```
from sklearn.model_selection import cross_val_score
from sklearn import metrics

docs
accuracy = cross_val_score(clf_DoS, X_DoS_test, Y_DoS_test, cv=10, scoring='accuracy')
print("Accuracy: %.5f (+/- %.5f)" % (accuracy.mean(), accuracy.std()) * 2)

precision = cross_val_score(clf_DoS, X_DoS_test, Y_DoS_test, cv=10, scoring='precision')
print("Precision: %.5f (+/- %.5f)" % (precision.mean(), precision.std()) * 2)

recall = cross_val_score(clf_DoS, X_DoS_test, Y_DoS_test, cv=10, scoring='recall')
print("Recall: %.5f (+/- %.5f)" % (recall.mean(), recall.std()) * 2)

f1 = cross_val_score(clf_DoS, X_DoS_test, Y_DoS_test, cv=10, scoring='f1')
print("f-measure: %.5f (+/- %.5f)" % (f1.mean(), f1.std()) * 2)
```

Figure 3. Confusion Matrix of the DoS and Probe attack
Figure 4. Confusion metrics of the DoS and Probe Attacks

```
# Create confusion matrix
Y_Probe_test2=clf_rfProbe.predict(X_Probe_test2)
# Predicted vs Actual labels
confusion = confusion_matrix(Y_Probe_test2, Y_Probe_test2)
```

Figure 5. U2R Attack

```
# Accuracy
accuracy = cross_val_score(clf_U2R, X_U2R_test, Y_U2R_test, cv=10, scoring='accuracy')
print("Accuracy: {} (+/-{})".format(accuracy.mean(), accuracy.std() * 2))
```

Figure 6. Denial of Service

```
# Cross-validation of U2R
confusion = cross_val_score(clf_U2R, X_U2R_test, Y_U2R_test, cv=10, scoring='precision')
```

Figure 7. Graphical representation of the U2R Attack Category
The Figure 6, Figure 7 and Figure 8 show the graphical representation of the DoS, U2R and Probe Attacks. Here, the number of features is plotted against the cross validation score. Also, the accuracy scoring is proportional to the number of correct classifications.

5. Discussion

Data analytics model for network intrusion detection presents a simple yet precise way of classifying and identifying attacks on the network. The model worked on two sets of dataset: train and test datasets. It divides the dataset into this two respectively and carries the testing and training phase on them. The performance metrics results achieved shows that the system performance in terms of accuracy, recall, f-measure score and precision is higher than those reviewed in previous researches. Also, this system is an improvement to existing systems since it can function well with small datasets and could detect attacks from both across and with the network.

Decision tree assists the network administrator to decide about the incoming traffic, i.e., whether the coming data is malicious or not by providing a model that separates malicious and non-malicious traffic. It allows taking less number of attributes and provides acceptable accuracy in reasonable account of time. From the results of the experiments, it is concluded that the system is more efficient with respect to finding attacks in the network with less number of features and it takes less time to construct the model. Also, the efficiency of the system has little or no regards for the size of the dataset and the number of features used to construct the decision tree.

6. Conclusion

The developed model was able to successfully classify and analysed the rate of network intrusion across and within the network. This was carried out using the various KDD processes. The dataset was pre-processed and cleaned using the data cleaning and data enriching techniques. The various attributes contained in the dataset was transformed so as to boast the accuracy of the result. In the course of this research, existing networks were examined. A classification model that classify attack label on the network was built. And each of the various intrusion types was evaluated based on the accuracy, precision, recall and f-measure score.

The evaluation of the model was based on the various attacks categories, the accuracy and precision were high for each of the categories. The model was able to successfully classify the various attack categories, check for each of their attributes and thus define their various rate.

References


