RESEARCH ARTICLE OPEN ACCESS

Intrusion Detection Based On Artificial Intelligence Technique

Shefali Singh[1], Dr. Zubair Khan[2], Krati Saxena[3]
M-Tech Research Scholar[1&3], Professor[2]
Department of Computer Science Engineering,
Invertis University, Bareilly-243123,
Lucknow, UP-India

ABSTRACT

Information Technology has become a main and important component to support critical infrastructure services in various sectors of our society. It is being used for sharing information and various operations. Many organizations are used to create complex network systems to give supply to the users. So due to this rapid expansion of computer usage, the security of the system has become very important. On every new day, a new kind of attack is being faced by the organizations. There were many methods which have being proposed for the development in the field of intrusion detection using artificial intelligence techniques. In this paper, we will have a look on a technique: K-Nearest Neighbors Algorithm for the understanding Intrusion Detection System. K-Nearest-Neighbors is one of the simplest and effective classification methods. This research shows the performance of the techniques on various test data sets. And the accuracy rate and the error rate are being calculated for these test data sets. And Confusion Matrix has been created for different Test Datasets for different values of K. Here, a model is being developed or the detection of anomaly attacks.

Keywords: - Intrusion Detection System, K-nearest neighbor algorithm, Confusion Matrix, Accuracy Rate and Error Rate.

I. INTRODUCTION

From last few years, Information is the most important and useful part of any organization. These organizations take measures to safeguard this information from the attackers. This rapid expansion of computer usage and internet in any organization has changed the world of computer and network by heaps and bounds [1] [2].

INTRUSION DETECTION SYSTEM is one of the most important security systems to detect intrusion in the distributed networking environment. It is a system for the identification of attacks in the network and takes corrective action to prevent them [3]. The main function of Intrusion Detection System is to protect the system, analyse the attack and predict the behaviours of users whether it is malicious or not [4]. The process of monitoring the events that are occurring in a computer network and analysing them for sign of intrusions is known as Intrusion Detection [5].There are two main approaches to design IDs. They are:-

1. Misuse based Ids

It sets up the attack behaviours based on known attacks behaviours. In it, the behaviours have being collected from attribute database. If the attack behaviours are same as in the database then it can safeguard it before the attacker destroys our system.

2. Anomaly based Ids

It is totally different from misuse detection. Intrusion Detection System monitors the event that occurs in the network and notifies it as an attack. The main advantage of anomaly detection is that it can identity unknown attacks. It is also called behaviours based detection. The drawback of anomaly based detection is that it have high false positive rate.

In the field of Intrusion detection, the algorithm is used in two stages: tolerance phase and testing phase. Output of the test data is generated in tolerance phase, and then used to detect known and unknown data in testing phase [3].

The tolerance phase is described as follows:

- Define the sample data
- Generate output for the test data and
- Match each test data with sample data. If it matches with any sample data, it is discarded; otherwise it is it is stored in a collection of data.

The testing phase is quiet similar to tolerance phase. Here, matching of test data is done with the sample data. If it matches then, it is unknown attack, otherwise it is known attack.

In this paper, Intrusion Detection system is implemented by using K-Nearest Neighbour technique. This system detect whether the activity is malicious or not. The K-Nearest neighbour is a supervised learning technique. Supervised learning algorithm is those in which correct answers are known and the information is used to train the network. This type of learning algorithm utilizes both input and output vectors. The input vectors are used to provide the starting data, and the output vectors are being used to compare with input vectors to determine some error. It is machine learning algorithm. It is a method for the classification of the objects based on the knowledge which it has learned from the training of data. The main challenge is to find out the Euclidean distance between the data. KNN is a very simple and effective technique to implement the system [7]. Our approach is to show the performance of the technique on various data sets and the accuracy rate and the

error rate is being calculated for the test data sets. And Confusion Matrix has been created for different Test Datasets for different values of K. A model is being developed or the detection of anomaly attacks.

II. RELATED WORK ON IDS

Tao Xu et.al [3] had presented an Intrusion detection approach inspired by biological memory cell. Here, they have proposed a system that detects the attacks in the system using biological memory cell. There result provides better performance than ordinary anomaly detection approaches with higher true positive rate and lower false positive rate.

M. Govindarajan et.al [8] presented intrusion detection system using K-Nearest neighbour algorithm. They have presented the effectiveness of K-nearest neighbour algorithm in intrusion detection.

The paper that is represented by Zhenghui Ma et.al is K-Nearest Neighbour with a novel similarity measure for Intrusion detection. This paper tells the simple n effective similarity definition within the nearest neighbours for the application. This similarity rule is having a fast computation and gains a satisfactory performance on the test data sets [7].

III. PROPOSED WORK

a). DATASETS AND FEATURES

CAIDA stands for Cooperative Association for Internet Data Analysis. CAIDA's Centre is at the San Diego Supercomputing Center (SDSC), an extension of the University of California at San Diego (UCSD) where it was established in 1997 by Dr. Kc Claffy and Tracie Monk. It is an organization which cooperatively undertaken with an interest in keeping basic Internet capacity and efficiency of its usage in line with increasing demand. The dataset is downloaded from the official site of CAIDA. It is the backscatter data of year 2008 which has being captured by the tool called Wireshark. The Backscatter-2008 is the newest data set captured for Dos attack, so this data set will be analyzed by us in the study. This data set contains information useful for examining denial-of-service attacks.

The features that are being used in the study are:-

DURATION –Length (number of seconds) of the connection.

PROTOCOL_TYPE -Type of the protocol, e.g. tcp, udp, etc.

SERVICE - Network service on the destination, e.g., http, telnet, etc.

SRC_BYTES -Number of data bytes from source to destination.

DST_BYTES -Number of data bytes from destination to source.

FLAG -Normal or error status of the connection.

b). K- NEAREST NEIGHBOUR ALGORITHM

This system detect whether the activity is malicious or not. The K-Nearest neighbor is a supervised learning technique. Supervised learning algorithm is those in which correct answers are known and the information is used to train the network. This type of learning algorithm utilizes both input and output vectors. The input vectors are used to provide the starting data, and the output vectors are being used to compare with input vectors to determine some error [8]. It is machine learning algorithm. It is a method for the classification of the objects based on the knowledge which it has learned from the training of data. In it, the results of the query is being classified on the bases of majority of K-nearest neighbors. The purpose of the algorithm is to classify the new entries of data based on attributes and training samples.

Algorithm:-

Step 1). Take a sample dataset of n columns and x rows named as S.In which $n-1^{th}$ columns are the input vector and n^{th} column is the output vector.

Step 2). Take a test dataset of $n-1$ attributes and y rows named as T.

Step 3). Find the Euclidean distance between every S and T by the help of formula

$$Eculidean\ Distance = \sqrt{\sum_{i=1}^{y}\sum_{j=1}^{x}\sum_{l=1}^{n-1}(S_{(j,l)}-T_{(i,l)})^2}$$

Step 4). Decide a random value of K is the no. of nearest neighbors.

Step 5). Then by the help of these minimum distance and Euclidean distance find out the n^{th} column of each.

Step 6). And find out the same output values. If the values are same then it a malicious attack; otherwise it is not.

After the detection of activities whether it is malicious or benign attack, the accuracy rate and the error rate of the data sets output is being calculated. And Confusion Matrix has been created for different Test Datasets for different values of K. Then, a model is being developed or the detection of anomaly attacks. The accuracy rate is showing that how many outputs of the data of the test dataset are same as the output of the data of different features of the training dataset. The error rate is showing that how many outputs of the data of the test dataset are not same as the output of the data of different features of the training dataset. Confusion matrix helps us understand how IDS performed (correct/incorrect classification)[11].

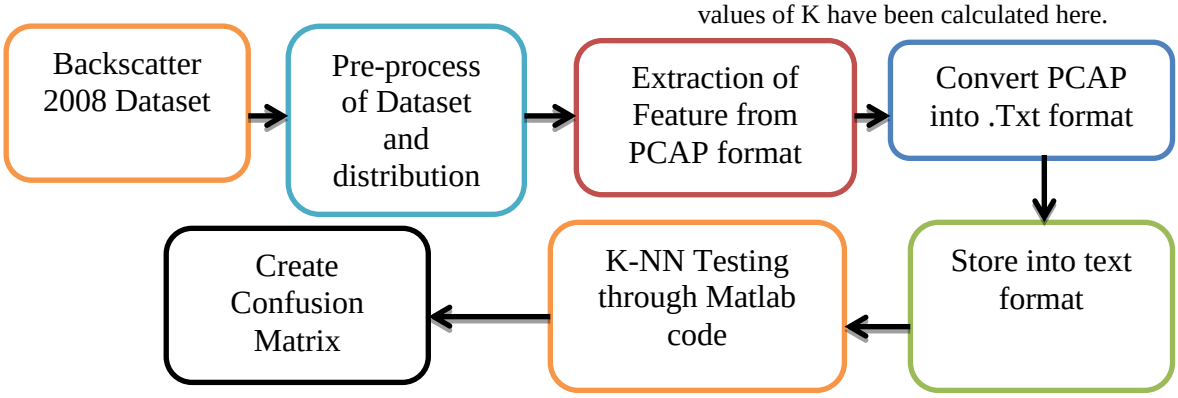
$$\text{True Negative Rate} = \frac{TN}{TN+FP} \times 100$$

$$\text{True Positive Rate} = \frac{TP}{TP+FN} \times 100$$

$$\text{False Negative Rate} = \frac{FN}{FN+TP} \times 100$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} \times 100$$

Proposed Model:-



- True Positive (TP) is when there is an intrusion in the events and the IDs detect it.
- True Negative (TN) is when there is no intrusion in the events and the IDs detect it.
- False Positive (FP) is when there is no intrusion and the IDs detect an intrusion in the event.
- False Negative (FN) is when there is an intrusion in the events and the IDs does not detect it [11].

IV. EXPERIMENTAL RESULTS

The data which is used is backscatter data of year 2008 which has being captured by the tool called Wireshark. The Backscatter-2008 is the newest data set captured for Dos attack, so this data set will be analysed by us in the study. This data set contains information useful for examining denial-of-service attacks. The features that are being used in the study are: - Duration, Protocol Type, Service, Source bytes, Destination bytes, and Flags. The K-Nearest Neighbour is the technique which is being used in training and tolerance phase of the dataset. The K-Nearest neighbour is a supervised learning technique. Supervised learning algorithm is those in which correct answers are known and the information is used to train the network. This type of learning algorithm utilizes both input and output vectors. The purpose of the algorithm is to classify the new entries of data based on attributes and training samples.

The accuracy rate and the error rate of the datasets are being calculated as a result. And Confusion Matrix has been created for different Test Datasets for different values of K. The accuracy rate is showing that how many outputs of the data of the test dataset are same as the output of the data of different features of the training dataset. The error rate is showing that how many outputs of the data of the test dataset are not same as the output of the data of different features of the training dataset. . Confusion Matrix is being made for the recognition of the intrusion. This matrix finds out the True Positive (TP) rate, True Negative (TN) rate, False Positive (FP) rate, and False Negative (FN) rate is the number of attacks incorrectly classified.

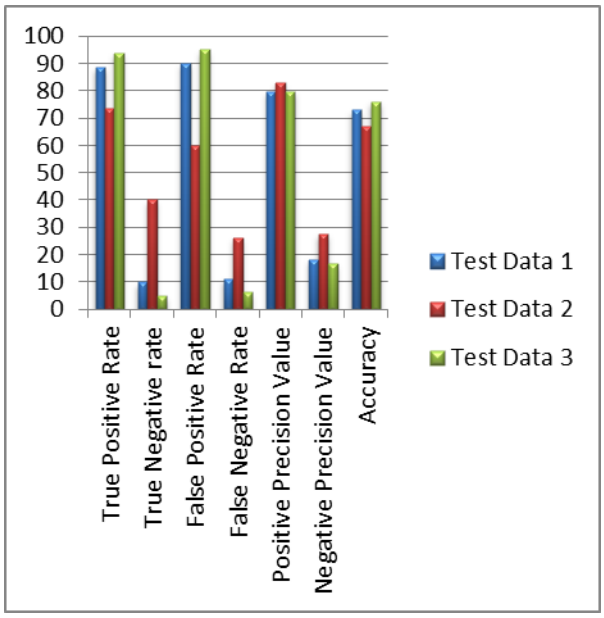
K means number of nearest neighbours. The results of different test data in form of confusion matrix for different values of K have been calculated here.

The results for the Test Data 1, Test Data 2, and Test Data 3 have been shown here for K=5.

	Test Data 1	Test Data 2	Test Data 3
True positive Rate	88.75%	75%	93.75%
True negative Rate	15%	40%	10%
False positive Rate	85%	60%	90%
False negative Rate	11.25%	25%	6.25%
Positive Precision Value	80.6818%	83.333%	80.6452%

Negative Precision Value	25%	28.5714%	28.5714%
Accuracy	74%	68%	77%

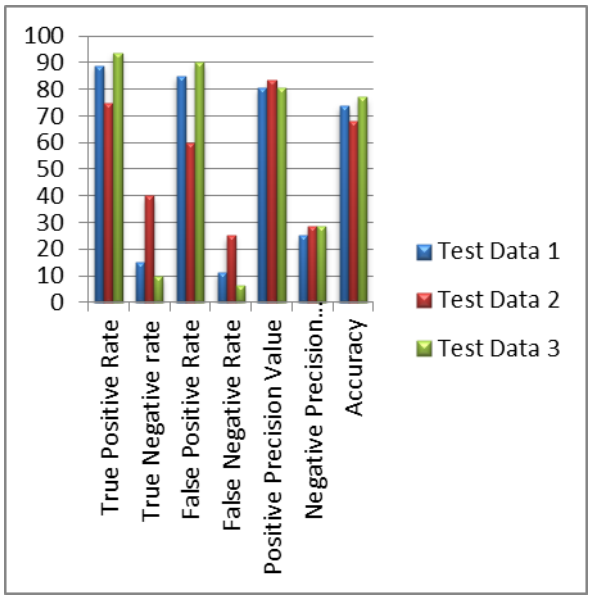
This is the Graph showing difference between Test Data 1, Test Data 2, and Test Data 3 for K=5.



The results for the Test Data 1, Test Data 2, and Test Data 3 have been shown here for K=7.

	Test Data 1	Test Data 2	Test Data 3
True positive Rate	88.75%	73.75%	93.75%
True negative Rate	10%	40%	5%
False positive Rate	90%	60%	95%
False negative Rate	11.25%	26.25%	6.25%
Positive Precision Value	79.7753%	83.0986%	79.7872%
Negative Precision Value	18.1818%	27.5862%	16.667%
Accuracy	73%	67%	76%

This is the Graph showing difference between Test Data 1, Test Data 2, and Test Data 3 for K=7.



V. CONCLUSION

The accuracy rate and the error rate of the datasets are being calculated as a result. The accuracy rate is showing that how many outputs of the data of the test dataset are same as the output of the data of different features of the training dataset. The error rate is showing that how many outputs of the data of the test dataset are not same as the output of the data of different features of the training dataset. And Confusion Matrix has been created for different Test Datasets for different values of K. The result concludes that when the number of K increases the accuracy rate increase and the error rate will decrease. The future work can be done using different technology and different datasets.

REFERENCES

[1] Damiano Bolzoni and Sandro Etalle : A 2-Teir Anomaly- based Network Intrusion Detection System, 2006, UK.
[2] D.A. Frincke and D. Tobin : A Frame work for cooperative Intrusion Detection System.
[3] Tao Xi : An Intrusion Detection approach inspired by biological memory cells, 2012, China.
[4] Rung Ching Chen and Kai-Fan Cheng : Using Rough Set and Support Vector machine for Network Intrusion Detection, 2009, Taiwan.
[5] Sandhya Peddabachigari and Johnson Thomas: Intrusion Detection System using Decision Tree and support vector machines, USA.

[6] Cheng-Leung Lui and Ting Yee Cheng : Agent based Network Intrusion Detection System using Data Mining approaches, 2005,ICITA.
[7] Zhenghui Ma and Ata Kaban : K-Nearest Neighbor with a Novel similarity measure for Intrusion Detection ,UK.
[8] M.Govindarajan and R.M. Chandrasekaran : Intrusion Detection using K-Nearest Neighbor, IEEE, 2009, India.
[9] P. S. Mann, 2004. Introduction to Statistics. 5th Edition. Printed in the United States of America. Johan Wiley & Sons. Inc
[10] Richard Jenson. A rough set aided system for sorting www.bookmarks.com
[11] Fawcelt, Tom (2006). "An Introduction to ROC Analysis". Pattern Recognition Letters 27
[12] CISCO System Ltd white paper. The science of Intrusion Detection System Attack Identification. World Wide Web, http://www.cisco.com/warp/public/cc/pd/sqsw/sqidsz/prodlit/idssa_wp.pdf [accessed 15 November 2011]
[13] Manish Kumar et.al ”intrusion detection system using Decision tree algorithm”,IEEE