An IDS by Correlation & KPCA with Neural Network Optimized By Genetic Algorithm

Harpreet Kaur
M. Tech. Scholar
Rayat Bahra University
Harpreetmultani6@gmail.com

Gaganpreet Kaur Bhalla
Assistant Professor, Rayat
Bahra University, Mohali
gaganb6@gmail.com

Abstract—An Intrusion Detection System is an application used for monitoring the network and protecting it from the intruder. Intrusion is a set of actions aimed to compromise these security goals. IDS have the computer security goals which are important for the data mining for extraction of data like confidentiality, integrity, and availability. This research study the performance measures of IDS is important for the security purposes. KDD 99 has 41 features. The IDS approached is used with the help of neural network technique of data mining and the Genetic algorithm is used as a classifier. The all features of KDD99 are used in this study. In this research, the feature is selected and extracted instead of using all features. The feature selection is done with the help of Correlation and feature extraction is done with the help of KPCA. The selection of features is according to the Eigen values of the features. The neural network is used for the change the weightage of the error. The neural network is basically runs many times and change weightage according to this.

Keywords—Data Mining, KDD Cup99, KPCA, Intrusion Detection System, Neural Network, Genetic Algorithm.

1. Introduction

An Introduction to Data Mining, Intrusion Detection System (IDS), and Neural Network (NN).

1.1. Data Mining

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the “knowledge discovery in databases” process, or KDD. The important data mining work is to extract the patterns of data from the huge amount of data that are present in any field like institutes, business etc. not from data itself. The data mining is the KDD process. For example, the data mining steps might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. There is no work for the individual working of the all processes of the KDD but all processes works together.

1.2. Intrusion Detection System

Internet and other web services are expanding very vastly in current years. Sometimes hackers or crackers misuse these services for the purposes or threat the data of that service or gaining some information for the security purposes. They harm the data of that web service and destroy it. So many techniques are established to protect the data like firewalls, antiviruses etc. but they could do it properly. So, the Intrusion detection system is used to protect the data from external forces that extract the data and harm it. Basically, we can say that IDS protects the computer security goals i.e. confidentiality, availability and integrity and IDS compromises according to these goals.
An ID basically is of two types: Anomaly and Misuse Intrusion detection system. In Misuse detection, attacks can be represented in the form of pattern or a signature in order to detect or prevent same attack in future. In anomaly detection category, deviation of normal usage behavior pattern is identified in order to correctly detect the intrusion.

How is data mining used in IDS? Pattern reorganization is the difficult task in the data mining. So, it can be handled by IDS and provide the classification is done by machine learning algorithms. Feature selection is the major problem in the learning system. So, Ids is working with KDD to perform in the less set of features and improve the accuracy and performance of the system.

1.3. Neural Network
Neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.

Neural network is made up of three layers i.e. input, hidden and output layers. A hidden layer is the processing layer. The hidden layer is in between the input and output layer.

Neural network on IDS is change the weightage of the input data according to the values.

This paper organized as follow: Section 2 present History, Section 3 also present KDD CUP 99 Dataset, Section 4 explain proposed approach, Section 5 consist experiments result, finally conclusion and future work is mentioned in Section 6.

II. LITERATURE REVIEW

The Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context. A simulation study was performed on KDD-99 cup data set for intrusion detection. Various classifier algorithms are applied over the data set and a learning model is tested [3]. The identified set of all features of KDD cup is used with SVM and neural networks [4].

Chena et al. [33] have used a Flexible Neural Tree (FNT) model for the IDS. The FNT model reduces the number of features. Using 41 features, the best accuracy for the Denial of Service (DoS) and User Gain Root (U2R) is given by the FNT model. The decision tree classifier supply the best accuracy for normal and probe classes, which are an improved less than the FNT classifiers.

A Genetic Algorithm (GA) approach with an improved initial population and selection operator, to efficiently detect various types of network intrusions. In the testing phase the Network Security Laboratory-Knowledge Discovery and Data Mining (NSL-KDD99) benchmark dataset has been used to detect the misuse activities. By combining the IDS with Genetic algorithm increases the performance of the detection rate of the Network Intrusion Detection Model and reduces the false positive rate [24].

The feature selection is done with the help of information gain and genetic algorithm is used for extracted that. The accuracy is checks according to the attacks of the KDD99 cup with intrusion detection [29].

3. KDD CUP 99 Data Set
KDD Cup’99 dataset used for benchmarking intrusion detection problem is used in our experiment. These are generated by processing the tcpdump segment of DARPA 1998 evaluation data set. This data set consists of 41 feature and separate feature (42nd feature) that labels the connection as ‘normal’ or a type of attack. The data set contains a total of 23 attack, these are grouped into 4 major categories:

3.1. Denial-of-Service (DoS)
In Denial-of-service attack, the attacker has the goal of limiting or denying service provided to the user, computer or network. Attacker tries to prevent genuine users from using a service. It is usually done by making the resources either too busy or too full and overflow.

3.2. Probing or Surveillance
Probing or Surveillance attacks have the main aim of gaining knowledge of the existence or configuration of a computer system or the network. The attacker then tries to harm or retrieve information about resources of the victim network.

3.3. User-to-Root (U2R)
User-to-root attack is attempts by an unauthorized user to gain administrative privileges. The attacker starts outs with access to a normal user account on the system (perhaps gained by sniffing password, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.
3.4 Remote-to-Local (R2L)
Remote-to-local attack is the kind of intrusion attack where the remote intruder consistently sends packets to a local machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.

III. PROPOSED METHODOLOGY

The KDD Cup 99 Data set is defined data set and automatically used in the system by using some software. It has 41 features that are also pre-defined in the system. Intrusion detection system is basically works on this data set.

KDD data set has two sets: Training set and testing set. Firstly, the feature extraction is done with the help of correlation and feature selection is done with the help of KPCA. These are the methods of data mining. Then, the ANN is used for processing the features to find the accuracy, precision and recall. At last, the genetic algorithm is used as a classifier and performed the error detection for the neural networks.

The flow chart of the proposed system is in figure 1. The steps of the proposed system flow chart are as follows:

1. The KDD99 cup is huge database. So, it is not possible to load huge database. So, the KDD99 is dividing into two sets: trained set and tested set.
2. The training set is loaded and for that the features selection is done for various dimensions separately.
3. The correlation is used for feature selection and selected features id defined in the class \{+1, -1\}.

![Figure 1: System flow diagram](image-url)
4. Then feature is extracted by KPCA. The feature extraction is done for the selected features. The feature extraction is done with the help of Eigen values and Eigen vectors.

5. Train neural network is used for input process for test set. The trained set generates input for the test set.

6. Neural network find some error for the input data and produce noise for test set.

7. The error is optimized by the genetic algorithm and if optimization is done properly no error in the input data, it goes to optimize weight.

8. All input data has some weight and that weight is optimized for error free data.

9. Again, the train neural is running and produces the test data set.

10. Test data is finally done with results and produce performance for the 10 fold values.

**IV. RESULTS AND DISCUSSIONS**

The performance of the proposed system is done with the help of Machine Learning - Confusion Matrix. Confusion matrix is a 2×2 matrix, where the rows represent actual classes; while the columns have the values correspond to the predicted classes.

![Confusion Matrix](image)

- TN is the number of true negative cases
- FP is the number of false positive cases
- FN is the number of false negative cases
- TP is the number of true positive cases

**Precision** = \( \frac{TP}{(TP+FP)\times 100} \)

**Recall** = \( \frac{TP}{(TP+FN)\times 100} \)

**Acc.** = \( \frac{TP+TN}{(TP+TN+FP+FN)\times 100} \)

**Experiment 1**

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>83</td>
<td>50</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>83</td>
<td>50</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
</tr>
<tr>
<td></td>
<td>59</td>
<td>50</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
</tr>
<tr>
<td></td>
<td>83</td>
<td>50</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
</tr>
</tbody>
</table>
Figure 4.1: Graphical Representation of the performance of 15 features

- Performance for 20 set of features:

Table 4.2: Performance for 20 set of features

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>83</td>
<td>50</td>
<td>83</td>
</tr>
<tr>
<td>83</td>
<td>50</td>
<td>83</td>
</tr>
<tr>
<td>83</td>
<td>50</td>
<td>83</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Figure 4.2: Graphical Representation of the performance of 20 features

- Performance for 25 set of features:

Table 4.3: Performance for 25 set of features

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>83</td>
<td>50</td>
<td>83</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
</tr>
<tr>
<td>83</td>
<td>50</td>
<td>83</td>
</tr>
<tr>
<td>90.5</td>
<td>50</td>
<td>90.5</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Performance for 30 set of features:

Table 4.4: Performance for 30 set of features

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>83</td>
<td>50</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>50</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>92.5</td>
<td>100</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
<td></td>
</tr>
<tr>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
<td></td>
</tr>
<tr>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
<td></td>
</tr>
<tr>
<td>92.5</td>
<td>50</td>
<td>92.5</td>
<td></td>
</tr>
<tr>
<td>91.9</td>
<td>50</td>
<td>92.5</td>
<td></td>
</tr>
</tbody>
</table>

These are the feature set values in tables and graphs as per the performance. Now, we combined the average value of the all the metrics of the data set and then perform the table and graph.

Table 4.5: Comparative Performance for different set of features

<table>
<thead>
<tr>
<th>Set of features</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>’15’</td>
<td>88.55</td>
<td>65</td>
<td>88.55</td>
</tr>
<tr>
<td>’20’</td>
<td>91.7</td>
<td>70</td>
<td>91.7</td>
</tr>
<tr>
<td>’25’</td>
<td>94.15</td>
<td>75</td>
<td>94.15</td>
</tr>
<tr>
<td></td>
<td>92.79</td>
<td>65</td>
<td>91.9</td>
</tr>
</tbody>
</table>
The graph shows the accuracy, precision and recall for the 4 different set of features. And according to this graph the accuracy is almost good for all the features and if we are not using the different attacks of the KDD99 cup the accuracy is also good without that. And the accuracy and recall is almost same because the recall is almost like the accuracy because it is true values of the functions.

5. CONCLUSIONS AND FUTURE SCOPE

CONCLUSIONS: This research study the performance measures of IDS is important for the security purposes. KDD 99 has 41 features. The IDS approached is used with the help of neural network technique of data mining and the Genetic algorithm is used as a classifier. The all features of KDD99 are used in this study. In this research, the feature is selected and extracted instead of using all features. The feature selection is done with the help of Correlation and feature extraction is done with the help of KPCA. The selection of features is according to the Eigen values of the features. The neural network is used for the change the weightage of the error. The neural network is basically runs many times and change weightage according to this. Then it creates the confusion matrix for calculation the value of precision and recall. The genetic algorithm is used as a classifier for error detection. The test results proves the feature selection of IDS get improved the selected features are used alone instead of the all features. All features are produces the redundancy, complexity in the system and decreases the accuracy. But a selected feature increases the accuracy, precision and recall for all the features set.

FUTURE SCOPE: The proposed models can be checked on different Intrusion detection system, to establish the new benchmarks on Intrusion detection. In future work, new hybrid model for intrusion detection can be built by optimizing the different machine learning algorithms. More parameters can be set for network features to improve the rate of intrusion detection, by further applying more techniques on proposed model this model can be laid as basic foundation for real life intrusion detection in future.

REFERENCES


