# NETWORK INTRUSION DETECTION SYSTEM USING REDUCED DIMENSIONALITY

M. Revathi [#1]

[#1]*Department Of Information Technology,*
Bharathiar University, Coimbatore, Tamilnadu, India

T.Ramesh[#2]

Assistant Professor
[#2]*Department Of Information Technology,*
Bharathiar University, Coimbatore, Tamilnadu, India

**Abstract**

Intrusion Detection System (IDS) is the science of detection of malicious activity on a computer network and the basic driver for network security. It is defined as a process of monitoring the events occurring in a computer system or network and analyzing them to differentiate between normal activities of the system and behaviors that can be classified as suspicious or intrusive. In this paper identifying the network attacks and comparing the performance of the algorithms is performed respectively. The Dimension Reduction focuses on using information obtained KDD Cup 99 data set for the selection of attributes to identify the type of attacks. The dimensionality reduction is performed on 41 attributes to 14 and 7 attributes based on Best First Search method and then apply the two classifying Algorithms ID3 and J48.

***Keywords:*** *Network Intrusion detection system; Best first search ; ID3; J48; Dimension Reduction; Cross Validation.*

## 1. Introduction

Network Security is the authorization of access to files and directories in a network. Users are assigned an identity number and password that allows them access to information and programs within their authority. Protecting a network from unwanted intruders. Intrusion Detection System (IDS) is the science of detection of malicious activity on a computer network and the basic driver for network security. Network security compromise the three security tokens such as confidentiality, integrity, availability. Confidentiality is the task of preventing unauthorized disclosure of information. Integrity the task of preventing unauthorized or accidental modification, creation or deletion of information. Availability the task of providing access to information and services when access is needed. If a system is able to assure that these three security tokens are fulfilled, it is considered secure.[2]

In this paper data mining classification algorithm is being used with the concept of Dimension Reduction. Dimension Reduction is applied using Best First Search which reduces the feature selection from 41 attributes to 14 and 7 potential attributes for classification. The proposed approach focuses on using information obtained KDD Cup 99 data set for the selection of attributes to identify the type of attack and then compares the performance of the ID3 with J48 by a randomly selected initial dataset with the reduced dimensionality. Furthermore, the results indicate that our approach provides more accurate results compared to the purely random one in a reasonable amount of time.

## 2. Data Set Description

The KDDcup99 Intrusion Detection datasets are based on the 1998 DARPA initiative, which provides designers of Intrusion Detection Systems (IDS) with a benchmark on which to evaluate different methodologies.[6]
*Attacks fall into one of four categories:*

- *Denial of Service (DoS):* Attacker tries to prevent legitimate users from using a service.
- *Remote to Local (R2L):* Attacker does not have an account on the victim machine, hence tries to gain access.
- *User to Root (U2R):* Attacker has local access to the victim machine and tries to gain super user privileges.
- *Probe:* Attacker tries to gain information about the target host.[4]

A set of data items, the dataset, is a very basic concept of machine learning. A dataset is roughly equivalent to a two-dimensional spreadsheet or database table. Here the 65000 records are selected as sample dataset out of 3, 11,029 Corrected KDD dataset connections. However, because the sample number of Probe, U2R, and R2L is less, the number of records of the attack types will be constant in any sample rate. The

remaining records out of 65,000 are 44,417 which are resulted by excluding the Probe, U2R and R2L types of records. Out of 44417, 20% of Normal connection is selected, and the Dos accounts remaining 80% of the dataset.

**3. Simulation Tool**

Simulation work for proposed Network Intrusion Detection System is done using WEKA[7].The Waikato Environment for Knowledge Analysis (WEKA) came about through the perceived need for a unified workbench that would allow researchers easy access to state-of the art techniques in machine learning. Weka provides three options:

- *Weka Explorer:* It has several panels that gives access to the main components of the workbench.
- *Weka Experimenter:* Allows user to create, run, modify and analyse the experiment in more convenient manner than when processing individually.
- *Weka Knowledge Flow:* It provides an alternative to the Explorer as a graphical front end to weka's core algorithm.

**4. Model Classification**

Here two algorithms ID3 and J48 are used for the classification of network attacks.

***4.1 J48 Decision Trees***

- J48 is an open source Java implementation of the C4.5 algorithm in the weka data mining tool[3].
- It builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy.
- At each node of the tree, J48 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other.
- Its criterion is the normalized information gain that results from choosing an attribute for splitting the data.
- The attribute with the highest normalized information gain is chosen to make the decision.
- For each attribute, the gain is calculated and the highest gain is used in the decision node.

***4.2 ID3***

- ID3, or Iterative Dichotomiser 3 Algorithm,[5] is a Decision Tree learning algorithm.
- Builds the tree from the top down, with no backtracking. Information Gain is used to select the most useful attribute for classification.
- ID3 is based on the Concept Learning System (CLS) algorithm. The basic CLS algorithm over a set of training instances C :

*Step 1:*

- ✓ If all instances in C are positive, then create YES node and halt.
- ✓ If all instances in C are negative, create a NO node and halt.
- ✓ Otherwise select a feature, F with values v1,.., vn and create a decision node.

*Step 2:*

- ✓ Partition the training instances in C into subsets C1, C2, ..., Cn according to the values of V.

*Step 3:*

- ✓ Apply the algorithm recursively to each of the sets Ci.

**5. Model Evaluation**

***5.1. Cross Validation***

Cross-validation, sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. 10-fold cross-validation is commonly used. In stratified K-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds.

### *5.2. Criteria for Evaluation*

To estimate the performance in the models Accuracy, Sensitivity, Specificity, and Receiver Operating Characteristics Curve (ROC) along with Kappa statistics and correctly classified Instance as criteria are employed. The accuracy, sensitivity and specificity were calculated by True Positive, False Positive, False Negative and True Negative.[1]

Accuracy means probability that the algorithms can correctly predict positive and negative examples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Sensitivity means probability that the algorithms can correctly predict positive examples.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

Specificity means probability that the algorithms can correctly predict negative examples

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

### *5.3. Confusion Matrix*

A confusion matrix is a visualization tool typically used in supervised learning (in unsupervised learning it is typically called a matching matrix). A confusion matrix that summarizes the number of instances predicted correctly or incorrectly by a classification model.

- The true positive rate (TPR) or sensitivity is defined as the fraction of positive examples predicted correctly by the model, i.e,
  TPR = TP / (TP + FN)
- The true negative rate (TNR) is defined as the fraction of negative examples predicted correctly by the model, ie,
  TNR = TN / (TN + FP)
- False positive rate (FPR) is defined as the fraction of negative examples predicted as a positive class the model, ie,
  FPR = FP / (TN + FP)
- The false negative rate (FNR) is the fraction of positive examples predicted as a negative class. i.e,
  FNR = FN / (TP + FN)

## 6. Performance Evaluation

Performance evaluation begins with the dimensionality reduction of original dataset which consist of 41 attributes and one class label. Using Best First Search it has been obtained the two set of potential dimensionalities 7 and 14 attributes. From the selected dimensionalities the experimental result shows that the performance of the reduced feature also predicts the classification in efficient manner. For experiments used Weka 3.6 data mining tool for analysing the results. The classification models can be evaluated using misclassification error rate and the area under ROC curve.

### *6.1. Dimensionality Reduction Algorithm*

Dimension Reduction techniques are proposed as a data pre-processing step. This process identifies a suitable low-dimensional representation of original data. Reducing the dimensionality improves the computational efficiency and accuracy of the data analysis.

*Steps :*

- ✓ Select the dataset.
- ✓ Perform discretization for pre-processing the data.
- ✓ Apply Best First Search algorithm to filter out redundant & super flows attributes.
- ✓ Using the redundant attributes apply classification algorithm and compare their performance.
- ✓ Identify the Best One.

The original dataset consist of 41 attributes and one class label. The following list out the attribute names

*(i) 41 Attributes:* duration, protocol type, service, Flag,, src_bytes, dst_bytes, land, wrong _ fragment, urgent,Hot,num_field_logins,logged_in,num_compromised,root_shell,su_attempted,num_root,num_file_creation,num_shells,num_access_files,num_outbounds_cmds,is_hist_login,is_guest_login,count, srv_count, serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_hosdst_same_srv_rate,dst_host_diff_srv_rate, dst_host_same _ src _ port _ rate, dst _ host _ srv _ diff _ host _ rate, dst _ host _serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate, dst _ host_srv_rerror_rate.,

Using Best First Search method we obtained two set of reduced dimensionalities. 7 potential attributes and 14 potential attributes which are listed in the table 2 and 3 respectively.

*(ii) 14 Attributes*: duration, service, flag, src_bytes, dst_bytes, count, srv _ count, serror_rate, rerror_rate, dst _ host _ same _ srv _ count, dst_host_srv_rate, dst _ host _ rerror _ rate , dst _ host _ diff _ srv_byte, dst_host_ same _ src _port_rate.

*(iii) 7 Attributes* : Protocol Type, Service,Srcbytes, Dstbytes,count, diff_srv_rate, dest_host_srv_count,

### *6.2. Best first Search*

Best First Search (BFS) uses classifier evaluation model to estimate the merits of attributes. The attributes with high merit value is considered as potential attributes and used for classification Searches the space of attribute subsets by augmenting with a backtracking facility. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions.

## 7. Simulation Result

The Receiver Operating Characteristic (ROC) curve is usually used to measure the performance of the classification method. Here the ROC curve is a graphical plot of sensitivity, specificity for the attributes.

Table 1. Sensitivity, Specificity And Accuracy Based On 41 Attribute Feature Selections

| | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| **ID3** | 98% | 100% | 99% |
| **J48** | 97.5% | 99.9% | 97% |

In Table 1 the Sensitivity, Specificity and Accuracy is calculated for the dimensionality 41 Attributes. The result shows that using ID3 for the 41 attributes the Sensitivity is 98% and Specificity is 100% and Accuracy is 99% respectively. In the case of J48 the sensitivity is 97.5%, Specificity is 99.9% and Accuracy is 97%. The result of above three table's shows that ID3 and J48 method had highest accuracy and sensitivity with 14 and 7 attributes. ID3 has highest specificity for all three dimensionalities. Thus ID3 algorithms performance is higher than J48.

Table 2. Sensitivity, Specificity And Accuracy Based On 14 Attribute Feature Selections

| | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| **ID3** | 100% | 98% | 99% |
| **J48** | 99.5% | 97.5% | 98% |

Here in Table 2 the Sensitivity, Specificity and Accuracy is calculated for dimensionality 14 Attributes. The result shows that using ID3 for the 14 attributes the Sensitivity is 100% and Specificity is 98% and Accuracy is 99% respectively. In the case of J48 the sensitivity is 99.5%, Specificity is 97.5% and Accuracy is 98%.

Here in Table 3 the Sensitivity, Specificity and Accuracy is calculated for dimensionality 7 Attributes using the values of True positive and False Negative obtained in classification with the ROC Table values.

Table 3. Sensitivity, Specificity And Accuracy Based On 7 Attribute Feature Selections

| | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| **ID3** | 100% | 99% | 99% |
| **J48** | 97% | 97% | 97% |

The result shows that using ID3 for the 7 attributes the Sensitivity is 100% and Specificity and Accuracy is 99% respectively. But in the case of J48 its 97% for all three parameters.

Roc Curve Result

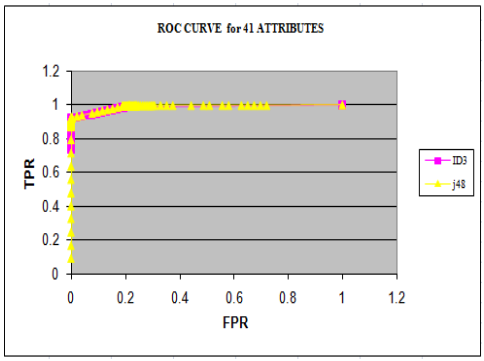


Fig 1. Graph for 41 Attributes

In this graph the parameters for the x – axis is the False Positive Rate (Specificity) and the Y – axis takes the True Positive Rate (Sensitivity) in fractions respectively. For 41 attributes the Sensitivity is 98% and specificity is 100% for ID3 and that for the J48 it is 97.5 % sensitivity and 99% specificity respectively.

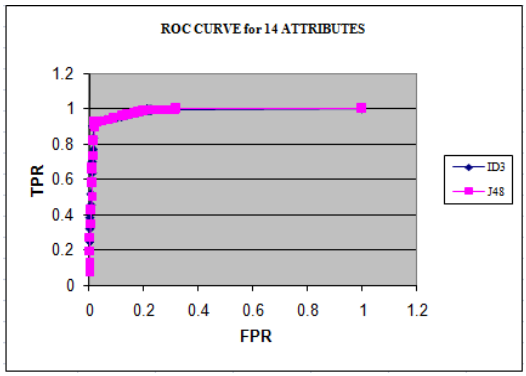


Fig 2. Graph for 14 Attributes

Here in Fig. 2 the graph is plotted for the 14 attribute dimensionality respectively. In this graph the parameters for the x – axis is the False Positive Rate (Specificity) and the Y – axis takes the True Positive Rate (Sensitivity) in fractions respectively.

For both algorithms the Sensitivity and Specificity value calculated previously is being used here. For ID3 it shows the Sensitivity and Specificity is 100 %( i.e., the fraction value 1) and for J48 the Sensitivity is 99.5% and the Specificity is 97.5% respectively.

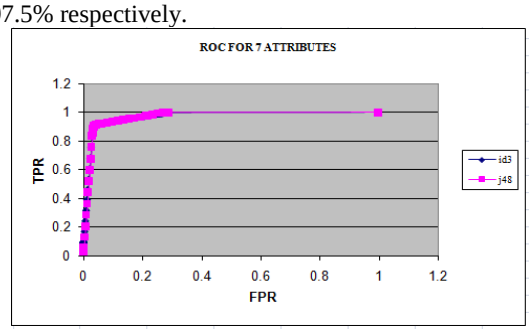


Fig 3. Graph for 7 Attributes

Here in Fig. 3 the graph is plotted for the 7 attribute dimensionality respectively. Compared with the previous 14 attribute here the sensitivity is 100 % and specificity is 99% for ID3and for J48 its sensitivity and specificity is 97% respectively. Thus the performance of ID3 here is more than that of J48.

Here from the above three graph it has been analyzed that the performance of the ID3 algorithm is relatively high in all the three dimensionalities with high accuracy than J48 algorithm. Thus it is concluded that ID3 performance is higher in this model.

## 8. Conclusion

The purpose of this work is to observe how these algorithm are used in the classification the Intrusion Detection Attacks. Thus the use of dimension reduction technique is an important task in this work to evaluate the performance. Here, two classification models such as ID3, J48, are used and compared their performance in three different dimensionalities using weka toolkit.

Using Dimensionality Reduction for three dimensionalities such as for 41 attributes 14 attributes and 7 attributes the classification of attacks are made and by applying the evaluation criteria the corresponding Specificity, Accuracy, Sensitivity are evaluated to get the respective True Positive, false positive rate for both the algorithms . The performance analysis of these algorithms is shown by the ROC CURVE. From the result it is observed that ID3 performs better classification and accuracy for three dimensionalities.

## References

[1] Anazida Zainal; Mohd Aizaini Maarof ; Siti Mariyam Shamsuddin,(2009): *Ensemble Classifiers for Network Intrusion Detection System*, Journal of Information , Universtiy Teknologi Malaysia.

[2] Andrea Janssen , (2009): *Hybrid Model* International Journal of Computer Science and Network Security, October 2009.

[3] Hossein, M. Shirazi, (2009): *Anomaly Intrusion Detection System Using Information Theory, K-NN and KMCAlgorithms* "Australian Journal of Basic and Applied Sciences.
[4] James, P. Anderson(1980): *Computer security threat monitoring and surveillance*, Washington, Pennsylvania, Journal of Computer Science and Network Security, USA,
[5] Kristopher Kendall (1999): *A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems* "Massachusetts Institute of Technology ,Journal of Computer Information technology..
[6] Li, Yuang, Guo, L.(2007): An *Active Learning Based TCM-KNN Algorithm for Supervised Network Intrusion Detection*. 26th Journal of Computers & Security.
[7] Sammany, M.; Medhat, T. (2007): *Dimensionality Reduction Using Rough Set Approach for Two Neural Networks-Based Applications*, Journal (RSE ISP) University, Tanta, Egypt.