An Efficient NIDS by using Hybrid Classifiers Decision Tree & Decision Rules

[1]Mrs. Nilofer Shoaib Khan, [2]Prof. Umesh Lilhore

M Tech Scholar, PG In charge
NIIST Bhopal (MP)

***ABSTRACT*: In the field of internet, network based application plays a vital role, where data transfers mostly in digital forms in various formats from source to destinations. In this digital exchange of information there are several possibilities of attacks and vulnerabilities. Intrusion detection systems are widely used to protect networks. An efficient detection of intrusion from network data set is a big problem which receives more attention from the various research communities. Various data mining classification techniques such as J-48 and SVM are widely used by researchers on various data sets such as KDD cup-99, NSL-KDD dataset. In this research paper we are presenting an efficient network intrusion detection system by using hybrid classifiers decision tree and decision rules for NSL-KDD dataset. An experimental study were performed by using weka-3-6 and MATLAB tool for existing J-48 and hybrid method and various performance improvement parameters such as precession, fmeasure, tprate and true positive rates are calculated.**

***Keywords*- Network based Intrusion detection system, KDD Cup-99, NSL-KDD, J-48, Hybrid decision tree**

1. INTRODUCTION

The advancement of Information Technology (IT) raised numerous security breaches. Therefore, to secure valuable resources over the public network, it is essential to implement an Intrusion Detection System (IDS). To defend against various cyber attacks and computer viruses, lots of computer security techniques have been intensively studied in the last decade, namely firewalls, anomaly and intrusion detection. Among them, Network Intrusion Detection (NID) has been considered to be one of the most promising methods for defending vibrant intrusion behaviors. Machine learning [3] is a burgeoning new technology for mining knowledge from data. In *data mining*, the data is stored electronically and the search is automated by computer.

Decision tree is one of the classifying and predicting data mining techniques, belonging to inductive learning and supervised knowledge mining. It can generate easy-to-interpret If-Then decision rule, it has become the most widely applied technique or methond among numerous classification methods IDS aimed to sort out various intrusive attempts on the computer network system based on the three important pillars of information security, i.e., confidentiality, integrity and availability of resources. In tree building stage, a decision tree algorithm can use its unique approach (function) to select the best attribute, so as to split training data set. The final situation of this stage will be that data contained in the split training subset belong to only one certain target class. Recursion and repetition upon attribute selecting and set splitting will fulfill the construction of decision tree root node and internal nodes. On the other hand, some special data in training data set may lead to improper branch on decision tree structure, which is called over-fitting.

Therefore, after building a decision tree, it has to be pruned to remove improper branches, so as to enhance decision tree model accuracy in predicting new data. It first gathers and analyze information from various sources within the computer network, triggers alarm to system administrators and blocks unauthorized access if an attack attempt is encountered. Various recent studies in IDS are evaluated based on a refined intrusion dataset with an error-free environment. However, the real network information deals with a huge amount of noisy data, and the IDS have to work in such an environment repeatedly.

2. RELATED WORK

The KDD Cup '99 dataset is the most well-known intrusion detection dataset available and researched by many researchers. The network traffic records in the dataset are classified as Normal or one of the four attack types i.e. denial of service, PROBE- network probe, R2L- remote to local and U2R- user to root attacks. In past various static machine learning algorithms have been evaluated and results are published. The results of the KDD'99 classifier learning contest, as summarized by Elkan [3], were all variants of the C5 decision tree algorithm (see Quinlan [4]).

After the contest a comprehensive set of other algorithms were tested on the KDD Cup 99 data, mostly with comparable results, were presented by Sabhani and Serpen [5], Sung and Mukkamala [6], Chavan, Shah et al. [7] and Peddabachigari, Abramham et al. [8]. The majority of results published are on the KDD Cup '99 `10%' training set only see Sung and Mukkamala [6], Kayacik, Zincir-Heywood et al. [9] and Lee, Shin et al. [10]. Some of the researchers extracted 11,982 records from KDD Cup 10% training dataset and build custom training datasets with 5,092 records and 6,890 test record see Chavan, Shah et al. [7], Chebrolu, Abraham et al. [11] and Chen, Abraham et al. [12].

Chavan, Shah et al. [7] Use a decision tree method for ranking of features per class. They reduced number of features from 41 to 13 for 'normal',16 for `probe', 14 for `dos', 15 for `u2r' and 17 for 'r2l' for experiment they evaluated it using artificial neural networks and fuzzy inference systems. Kayacik, Zincir-Heywood et al. [9] investigated the relevance of each feature provided by the KDD Cup '99 intrusion detection dataset in terms of information gain and presented the most relevant feature for each individual attack. Another important result was that nine features do not make any contribution for intrusion detection.

This novelty detection approach was employed to detect attack categories in the KDD dataset. The technique has achieved the detection rate of 96.71% of DoS, 99.17% of Probe, 93.57% of U2R and 31.17% of R2L respectively. However, due to the fact that no FP was reported by the research scientists a nearly impossible detection rate [5] of 93.57% of U2R category. Tavallaee et al. [15] described the importance of each feature in KDD '99 intrusion detection dataset for detection of DOS, PROBE, U2R L2R and Normal class.

They also discuss various problems of KDD Cup '99 datasets and created a revised version of the datasets, called NSL-KDD to address the some of known issues. They modified the class distributions by cleaning the training and testing datasets. This will avoid biasness towards the more frequent records. Ben Amor et al. [16] performed comparative analysis of decision tree vs naive bayes and found that decision tree performs slightly better than naive bayes. They also found that building naive bayes computational model is faster than of decision tree. Decision trees generally have very high speed of operation and high attack detection accuracy.

3. EXISTING CLASSIFIER METHODS FOR NIDS & CHALLENGES

3.1 J48- It utilizes a divide-and-conquer approach and recursively create a decision tree based on the greedy algorithm11. It consists of the root node, branches, parent nodes, child nodes and leaf nodes. A node in a tree denotes dataset attributes; every child node derives labeled branches about the possibilities of attribute values from the corresponding node called parent node.

3.2 Negative Selection Method-The basic approach of Negative Selection Algorithm is anomaly detection using negative detectors and it was originally introduced by Forest based on Clonal detection process in the immune system to prevent from auto immunity. The NSA initiates random detectors and separate the one that perceive self patterns, and detects non self pattern which is resulted by the collection of detectors.

4. PROPOSED METHOD

There are many existing mechanisms for Intrusion detection system, but the major issue is the security and accuracy of the system. To improve the problem of accuracy and the efficiency of the system a very common classification approach i.e. decision tree is used.

In this research paper we are presenting an efficient network intrusion detection system by using hybrid classifiers decision tree and decision rules for NSL-KDD dataset. In this framework NSL-KDD dataset is given to Pre-processing stages which classify in J-48 algorithm with negative selection method and reduce irreverent features from the data set so that data with less number of features will require feed to the classifier and will provide efficiency to the classifier. Machine learning tools WEKA are used to analyse the performance of datasets.

4.1 Proposed Algorithm- Proposed Hybrid classifiers decision tree and decision rules For NIDS

```
Step-1 Select network data set
Where D ← Stored data from database, N ← all feature set, th← threshold value
Step-2 for i = 1 → n do
Step-3 For (Testing datai ∈ Testing data) Testing data i Class := Self
3.1 F = F - Fi
3.2 ac = calculateAccuracy(F)
3.4  if ac ≤ th then
3.5 break
3.6 end if
3.7 For (Testing datai ∈ Testing data) Testing data i Class: = Self
3.8 For (Detector j ∈ Detector repertoire ) If (Matches Testing datai , Detector j ) Testing data i Class := Non Self
3.9 Break
3.10 End If
3.11 End For
3.12 End For
Step-4 Apply learning algorithms
Step-5 Classified Data
```

5. IMPLEMENTATION & RESULT ANALYSIS

All the experiments in this paper are implemented using WEKA 3.7.9 machine learning tool and MATLAB 2013a.

5.1 Dataset Description -The proposed system is evaluated using publicly available NSL-KDD intrusion detection dataset which is an enhanced version of the KDD99 intrusion detection dataset. KDD99 dataset is the only well-known and publicly available data set in the area of intrusion detection [14]. It is still widely used in evaluating the performance of proposed intrusion detection algorithms. On the KDD99 intrusion detection dataset 78% of training instances and 75% of test instances are duplicated. Hence the NSL-KDD dataset is generated by removing redundant instances in both the training and test data of the KDD99 intrusion detection dataset [12]. This dataset has 41 features and one class attribute. The training data contains 24 types of attacks and the testing data contains extra 14 types of attacks. The attacks in this dataset are categorized in one of the four attack categories (DoS, Probing, User to Root and Remote to Local attacks)

Though NSL-KDD dataset is enhanced version of the KDD99 dataset we observed two basic problems in this dataset. That is some records have same value for all the 41 features, however they are labeled to different classes (one as normal and the other as attack). The second observation we made is there is a feature called num_outbounds_cmds which has a value of zero for all the records in both the training and testing data. This feature will not have any contribution in identifying attacks from normal profiles. Hence we made two improvements in using NSL-KDD dataset: we removed all ambiguous records and the num_outbounds_cmds feature from the dataset.

5.2 Data Pre-processing-After calculating information gain for each of the features in the training data, for the J48 classifier we selected 20 features by applying the optimal feature selection algorithm with T=0.9. The J48 classifier is built using the selected features and the KDDTrain+ full training data. For the ensemble one-class J-48 classifiers 11 features are selected from the 20 features by applying optimal feature selection algorithm with T=0.9.

5.3 Confusion Matrix- The confusion matrix was used to evaluate the performance of the IDS. A confusion matrix is a specific table layout that allows visualization of the performance of IDS. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). In the binary class IDS, the intrusion detection system is mainly discriminate between to classes, "Attack" class (malicious threats or abnormal data) and "Normal" class (normal data). Table 1 shows the confusion matrix.

		Predicted		**Total**
		Normal	**Attacks**	
Actual	**Normal**	**TN**	**FP**	**TN+FP**
	Attacks	FN	TP	FN+TP
Total		TN+FN	FP+TP	

Table

5.3 Confusion Matrix

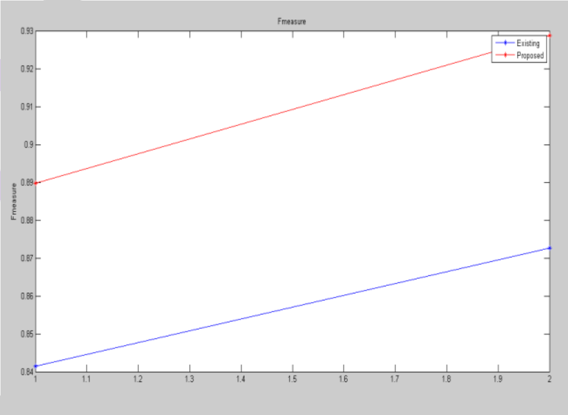
True Positives (TP): The number of attack classified as attack.

True Negatives (TN): The number normal classified as normal.

False Positives (FP): The number of normal classified as attack.

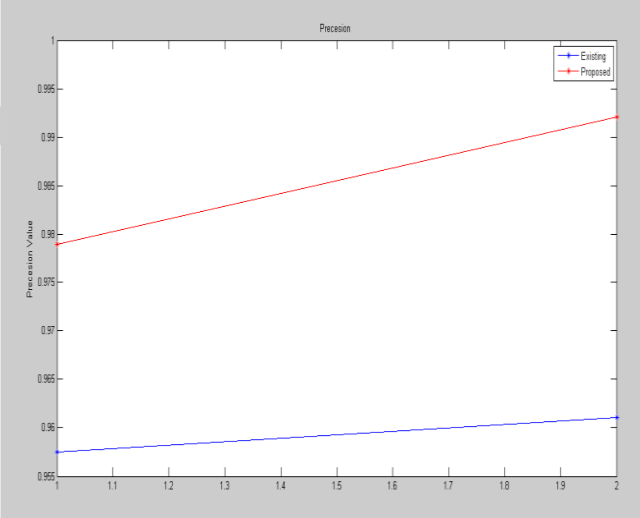
False Negatives (FN): The number of attack classified as normal

5.1 Precision- Precision is a description of random errors, a measure of statistical variability.



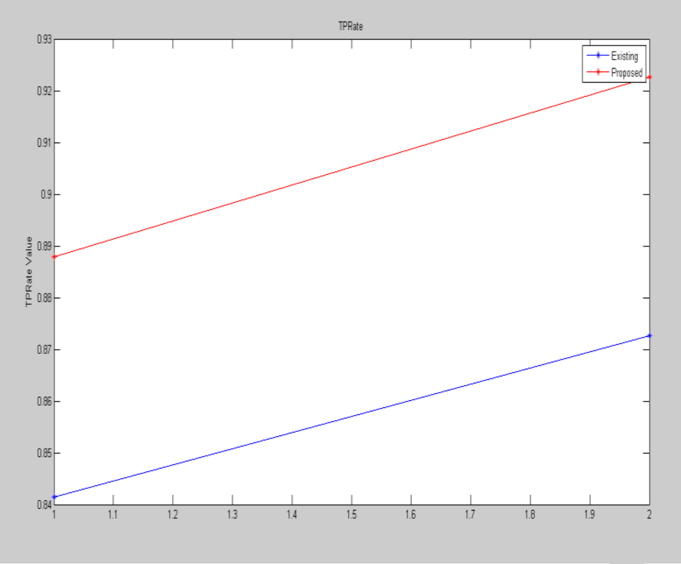
5.2 True Positive Rate-It measures the proportion of positives that are correctly identified

TPR = TP/P = TP / (TP+FN) Where TP = True Positive, P = Positive, FN = False Negative



5.3 F measure-It is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score.

FMEASURE = (2* PRECISION* RECALL) / (PRECISION + RECALL)



6. Conclusions & Future Work

Some of the existing researches have also been discussed in this paper like some types of pre-processing approaches such as data mining, neural network models, artificial intelligence, have been examined for achieving the better rate detection within the intrusion-detection-system. In this research paper proposed method hybrid classifiers decision tree and decision rules for NSL-KDD dataset performs outstanding in terms of precision, true positive rate. In future work we can implement proposed method with more realistic data sets such as Kyoto 2006, 2009, 2012 and more classifiers can be used for comparisons.

REFERENCES

[1] Jamal Hussain, Samuel Lalmuanawma," Feature analysis, evaluation and comparisons of classification algorithms based on noisy intrusion dataset", 2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016), ScienceDirect, Procedia Computer Science 92 (2016) 188 – 198.

[2] Preeti Aggarwal, Sudhir Kumar Sharma," Analysis of KDD Dataset Attributes - Class wise For Intrusion Detection", 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), ScienceDirect, Procedia Computer Science 57 (2015) 842 – 851

[3] Kumar, Vipin, Jaideep Srivastava, and Aleksandar Lazarevic, "Managing cyber threats: Issues, approaches, and challenges". Vol. 5.Springer,2006.

[4] Maheshkumar Sabhnani and Gursel Serpen, "Why Machine Learning Algorithms Fail in Misuse Detection on KDD Intrusion Detection Data Set". ACM Transactions on Intelligent Data Analysis,(pp.403-415) (2004).

[5] M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier".Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), pp. 172– 179, 2003.

[6] J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory". ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262–294, 2000.

[7] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed analysis of the KDD CUP 99 Data Set". In the Proc. Of the IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), pp. 1-6, 2009.

[8] S. Revathi, Dr. A. Malathi, "A detailed analysis of KDD cup99 Dataset for IDS". International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 12, December – 2013.

[9] R. P. Lippmann, D. J. Fried, and I. Graf, "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation". In Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX'00), (2000).

[10] "Nsl-kdd data set for network-based intrusion detection systems". Available on: http://nsl.cs.unb.ca/NSL-KDD/, November 2014.

[11] Lee W., and Stolfo S.J., "A framework for constructing features and models for intrusion detection systems". ACM Transactions on Information and System Security, 3 (4) (pp. 227-261) (2000).

[12] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, " Top Ten Data Mining Algorithms". Knowledge and Information Systems Journal, Springer-Verlag London, vol. 14, Issue 1, pp. 1-37, 2007.

[13] Lei Li, De-Zhang Yang, Fang-Cheng Shen, "A Novel Rule-based Intrusion detection System Using Data Mining". In the Proc. Of 3rd IEEE International Conference on Computer Science and Information Technology, pp. 169-172, 2010.

[14] J. E. Gaffney and J. W. Ulvila, "Evaluation of intrusion detectors: A decision theory approach". In Proceedings of the 2001 IEEE symposium on Security and Privacy, pages 5061, Oakland, CA, USA, 2001.