I. J. Computer Network and Information Security, 2013, 11, 51-57
Published Online September 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2013.11.07

A New Model for Intrusion Detection based on Reduced Error Pruning Technique

Mradul Dhakar, Akhilesh Tiwari
Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior (M.P.), India
mraduliitm@gmail.com, atiwari.mits@gmail.com

***Abstract* —** The increasing counterfeit of the internet usage has raised concerns of the security agencies to work very hard in order to diminish the presence of the abnormal users from the web. The motive of these illicit users (called intruders) is to harm the system or the network either by gaining access to the system or prohibiting genuine users to access the resources. Hence in order to tackle the abnormalities Intrusion Detection System (IDS) with Data Mining has evolved as the most demanding approach. On the one end IDS aims to detect the intrusions by monitoring a given environment while on the other end Data Mining allows mining of these intrusions hidden among genuine users. In this regard, IDS with Data Mining has been through several revisions in consideration to meet the current requirements with efficient detection of intrusions. Also several models have been proposed for enhancing the system performance. In context to improved performance, the paper presents a new model for intrusion detection. This improved model, named as REP (Reduced Error Pruning) based Intrusion Detection Model results in higher accuracy along with the increased number of correctly classified instances.

***Index Terms* —** Data mining, intrusion detection, REP, K2, KDDCup'99

I. Introduction

The valuable importance of Intrusion Detection System (IDS) has proven its significance in eliminating the leading security issues. IDS, is a security system or software with a goal of effectively detecting intrusions in a surveillance environment. IDS monitor the system activities within the environment and meanwhile detect for intrusions. In order to ban the access and preserve the data from intrusions, IDS is widely preferred in the current scenario.

IDS, has been continuously revised in order to improve the detection process with increased performance. In this regard, it has been incorporated with some of the dominating research fields such as statistics, neural network, data mining etc. When applying data mining technology to intrusion detection systems, it can mine the features of new and unknown attacks well, which is a maximal help to the dynamic defense of intrusion detection system [1]. Thus, making the approach a preferable alternative along with the well suited tools and techniques defined under data mining.

Moreover up-gradations are done regularly in techniques to develop a more efficient system and adding in intelligence to precisely detect the known as well as unknown attacks. With this view in mind the paper is an effort which considers the objective of making the IDS more efficient and intelligent.

The purpose of this paper is to propose a new classification model for intrusion detection named as Reduced Error Pruning based IDS model. The model is able to sufficiently classify the attacks with the raised performance rate. The proposed model is compared against typically preferred K2-based IDS model for performance evaluation. The experimental assessment proves proposed model highly feasible for intrusion detection with desirable accuracy and precise classification of intrusions. Also the REP-based IDS model is analyzed in aspect of performance evaluation with the available Data Mining techniques for intrusion detection where the maximal accuracy is achieved by the proposed model.

The paper is further organized as follows: section II describes about IDS and Data Mining along with its techniques, types and sorts of attacks detected; section III explains the methodology for the proposed model; section IV elaborates the proposed model; section V explores the Experimental Analysis and lastly section VI ends up the paper with a conclusion.

II. Intrusion Detection System And Data Mining

The Intrusion Detection System is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents [2].The strength of IDS is its ability to precisely detect the various forms of abnormalities within the network traffic. It can be considered as a combination of hardware and software where the system scans for intrusions.

Intrusion Detection System is an important detection approach used as a countermeasure to preserve data integrity and system availability from attacks [3]. The IDS continuously seeks for intrusions in a monitoring environment by tracking the network activities. Each time when IDS suspects intrusions, either by performing a match from knowledge gained from previous attempted

intrusions or by estimating the deviation of normal activity, an alarm is set indicating the presence of intrusion.

The process of detecting intrusion using Data Mining (shown in Fig. 1) can be in general understood as:

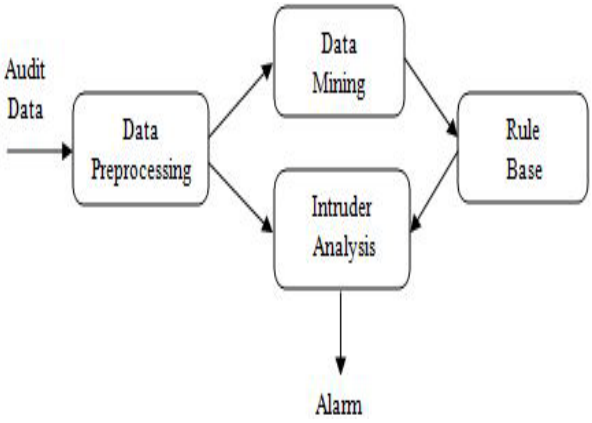

Figure 1: Working of IDS

Although IDS has justified its purpose to a higher extent, still there is a need for advancement in order to meet the current requirements. Also an aim to reduce the misleading detection of normal activities (known as false alarm) has been sustaining till date.

With this concern, Data Mining has become a favorable incorporated domain that made IDS to work progressively to tackle the shortcoming. Although data mining has become a very useful technique by reducing the information overload and improving the performance of the IDS [4], IDS with Data Mining have made many appreciable efforts for detecting the anomaly more precisely. It is difficult to remove all possible errors due to the enormous variety and complexity of today's networks. Data Mining along with its various applicable techniques has made the IDS cope up with the demanding trends making the system more efficient and satisfactorily intelligent.

A. IDS Techniques

Intrusion detection techniques can be categorized into misuse detection, which uses patterns of well known attacks or weak spots of the system to identify intrusions; and anomaly detection, which tries to determine whether deviations from the established normal usage patterns can be flagged as intrusions [5]. They are further elaborated as:

Misuse Detection: Also known as Signature-based Detection is the approach where the known attack patterns are pre-stored. These patterns serve as a signature for the type of intrusion to be detected by the IDS. Whenever a match to the signature is found, an alarm is set pointing to the presence of an intrusive activity. Since it works on the basis of predefined signatures, it is unable to detect new or previously unknown intrusions.

Anomaly Detection: In anomaly detection, profiles for normal behavior are defined which can be used to detect the abnormality within the system or network. The abnormality is tracked in the perspective of deviation of user activity from that of normal profiles. A threshold value (or a limiting factor) is assigned such that whenever the value of deviation goes beyond the threshold, an alarm is set. However anomaly is able to detect new intrusion but the compulsion for limiting factor involvement results in a high percentage of false positive rates.

B. Types of IDS:

Based on the resources being monitored by IDS, it is categorized into two types, Host-based Intrusion Detection System and Network-based Intrusion Detection System.

Host-based Intrusion Detection System (HIDS): The HIDS is a software application which is installed onto a system in order to protect it from intrusion. The HIDS continuously monitors the event logs and file attributes generated by various applications and operating system. HIDS is considered good in detecting buffer overflow attacks however it is OS dependent and require some preplanning before implementation.

Network-based Intrusion Detection System (NIDS): NIDS in comparison to HIDS is installed in a network in order to detect intrusions. NIDS analyzes packets that flow in the network. NIDS is considered good in providing security against the DoS type of attacks also OS independent and easy to deploy.

C. Sorts of Attacks detected by IDS:

Following are the basic four categories of attacks being detected by IDS:

Denials-of-Service (DoS): Denial of Service attacks prevent the legitimate users from accessing the services of a host or network resources. This is usually done by making the resources either too busy or overflow, which as a consequence results in the rejection of services requested by the legitimate users.

Probing or Surveillance: Probing or Surveillance attacks, as the name implies, are the attacks performed by probing or gaining knowledge about the network configuration and its vulnerabilities with a motive to harm or retrieve information about the resources of the victim network.

User-to-Root (U2R): User-to-Root attacks are attempts by a non-privileged user to gain administrative privileges i.e. a local user who is the intruder tries to gain access to a computer system as a root. By gaining the root access, the aim of the intruder is to compromise the vulnerabilities of the system.

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 with the intention of gaining access as a local user. The intruder tries to explore the vulnerabilities by exploiting the acquired privileges of a local user.

III. Methodology

This section involves discussion of the two algorithms of data mining classification approaches, K2 and REP. The K2 algorithm (under BayesNet) in literature has been a widely used algorithm due to its effective classification. While the REP algorithm (under Decision Tree) in contrast to K2 not only provides effective classification results but also involves the pruning of the tree with fast decision learning capability.

A. K2 Algorithm:

K2 is one of the generally used algorithms under Bayesian Networks. It is an algorithm for constructing BayesNet from a database of records. It is a heuristic search algorithm that searches for intrusions on the basis of a predefined ordered set of instances. The algorithm requires a set of nodes, a previously known order of the nodes, an upper bound on the number of parents a node may have, and a database containing possible cases. It starts by assuming that a node has no parents, after which, in every step it adds incrementally the parent whose addition mostly increases the probability of the resulting structure [6]. The algorithm terminates adding parents to the nodes, when the adding of single parent cannot increment the probability of the network given the data.

Though the K2 algorithm works well for intrusion detection along with the satisfactory error rate, it however requires an ordered set defined as heuristic and it also lacks in computational simplicity.

B. Reduced Error Pruning (REP)Algorithm:

REP is a fast decision tree learner classifier of data mining technique. It uses a validation data set for estimating generalization error. The error is pruned for each node in the tree. Essentially the node with the highest reduced error rate is pruned.

Pruning can be understood as a technique whose objective is to reduce the size of the decision tree by removing parts of a tree that allow better classification of instances. Pruning therefore reduces the classification complexity with increased accuracy. The accuracy is improved by the reduction of over-fitting and by removal of the parts of the tree classifier that may be based on noisy or erroneous data. Hence, a pruning of a tree is a sub-tree of the original tree with just zero, one or more internal nodes changed into leaves [7].

The pruning in REP is always done at leaves where each node is replaced with its most popular class. First, the training data are split into two subsets: a growing set (usually 2/3) and a pruning set (1/3) [8]. The growing phase is used to grow the rules for constructing classification tree while pruning phase performs pruning.

Next, an error rate is calculated for each node, estimated as the number of instances that are misclassified on a validation (pruning) set by propagating errors upward from the leaf nodes. The difference in error rate is determined by replacing the most common class resulting from a node. If the difference is a reduction in error then the sub-tree below the node can be considered for pruning. The leading benefits of this classification technique are its simplicity and the speed of decision learning. However the tree needs large amount of data for pruning, REP tree is still assumed as a more accurate classification tree.

In reference to the above working discussion of REP, following are the algorithmic steps undertaken.

Algorithm1: Reduced Error Pruning Algorithm

1 Select dataset
2 Split the input dataset into two subsets, a growing set and a validation set.
3 Repeat the pruning phase i.e. step 4 and 5 for every node in the tree
4 Evaluate the impact on the validation set i.e. error rate for each node.
5 Remove the node which maximally improves the accuracy of the validation set i.e. the node with highest reduced error rate.

IV. Proposed Model

The proposed model i.e. REP based Intrusion Detection Model is an enhancement to the other IDS models. The proposed model incorporates a reduced error pruning approach which aims at pruning the classification tree by reducing error rate for the system. The model shown in Fig. 2, involves the preprocessing of the KDD dataset, splitting it into train-test sets, applying of the REP algorithm (consisting of tree construction and pruning), then evaluating the performance for classification, and finally visualizing the outcomes accordingly.

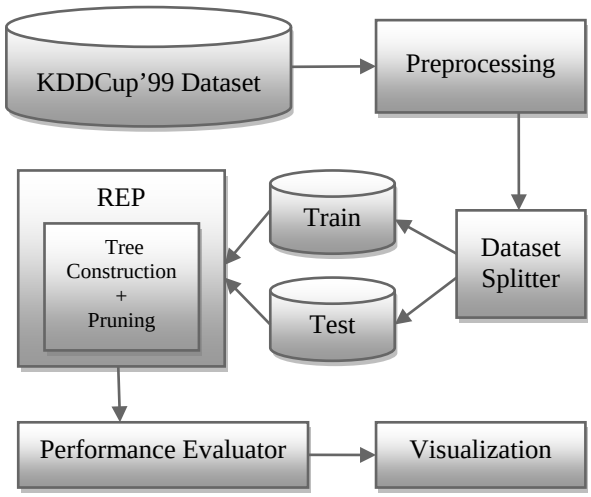

Figure 2: REP-based Intrusion Detection Model

A. Description of the Proposed Model

KDDCup'99 Dataset: The KDDCup'99 dataset is the most widely used dataset since 1999. It has become an extensively preferred dataset used by researchers to evaluate the effectiveness of IDS models. The KDDCup'99 dataset was originated from processing the tcpdump segment of DARPA 1998 evaluation dataset. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a

military network environment [9].The data set consists of 41 features and a separate feature (42nd feature) that labels the connection as 'normal' or a type of attack. Since the KDDCup'99 data set consists of a huge number of data records (about 5 million) for the training the intrusion detection system, difficulty is faced in analyzing the dataset as a whole. Henceforth The 10% of KDDCup'99 dataset has been chosen from the original KDDCup'99 dataset with 494021 records to train and test the proposed model.

Pre-processing Phase: In the pre-processing phase, the 10% of KDD dataset has been modified in order to make the classification easier. The dataset here is pre-processed by categorizing the prediction class into 4 different types of attacks i.e. probe, dos, u2r and r2l. This categorization hence would help in making the evaluation simpler.

Dataset Splitter: In the dataset splitting phase, the KDDCup'99 dataset splits into two parts: one is training set and the other is test set. The train set and test set is divided in the ratio of 3:2 (i.e. the train set is the 66% and test set is the 44% of KDDCup'99 dataset) by the splitter. The splitter opt instances randomly for the training of the model from the original dataset while the rest is available for testing of the trained model. Therefore KDDCup'99 dataset used in the model have 494021 instances which get partitioned randomly into 326054 instances for training while remaining 167967 instances for testing.

Performance Evaluator: The Performance evaluator phase assesses the performance of REP based IDS model by calculating the following parameters:

a) True Positive Rate (TPR):

$$TPR = \frac{TP}{TP + FN}$$

b) False Positive Rate (FPR):

$$FPR = \frac{FP}{TN + FP}$$

Where TP (True Positive), FN (False Negative), FP (False Positive) and TN (True Negative) can be defined as follows [10]:

- True Negative (TN): The percentage of valid records that are correctly classified.
- True Positive (TP): The percentage of attack records that are correctly classified.
- False Positive (FP): The percentage of records that were incorrectly classified as attacks whereas in fact they are valid activities.
- False Negative (FN): The percentage of records that were incorrectly classified as valid activities whereas in fact they are attacks.

These parameters described above can also be illustrated through Table I.

TABLE I CONFUSION MATRIX OF TN, TP, FN AND FP

	Correctly Classified	Incorrectly Classified
Valid Record	True Negative (TN)	False Positive (FP)
Attack Record	True Positive (TP)	False Negative (FN)

Confusion Matrix is one of the other parameters in literature to analyze the performance of the model. A confusion matrix is a tabular visualization of the performance of an algorithm. The column in the matrix represents the instances of a prediction class while the row represents the instances of an actual class.

Visualization: In this phase the performance results of the REP based IDS model can be visualized through various means such as text, graph etc. On the basis of the results obtained in this phase, the effectiveness of the model can be examined.

V. EXPERIMENTAL ANALYSIS

This section contains the experimental results obtained from REP based IDS model along with its comparison with K2 based IDS approach. The two algorithms when compared, it has been observed that more satisfactory results are obtained by applying REP. Since both algorithms consequently results in the classification of the data as normal or the type of attack accordingly, the time taken for evaluating the data is nominal in case of REP. Also, the accuracy obtained for REP is raised appreciably than K2. Hence it would not be irrelevant to say that REP has proved itself to be more preferable classification technique than K2.

Table II shows the comparison of TPR & FPR between REP and K2

TABLE II CLASS WISE PERFORMANCE OF REP AND K2

Class	REP		K2	
	TPR	*FPR*	*TPR*	*FPR*
DoS	1.000	0.001	0.988	0.000
Probe	0.978	0.000	0.978	0.005
R2L	0.982	0.000	0.959	0.001
U2R	0.667	0.000	0.810	0.005
Normal	0.999	0.000	0985	0.002

Table III and IV contain the confusion matrices of the attacks classified by REP and K2 respectively.

TABLE III CONFUSION MATRIX OF REP

	DoS	Probe	R2L	U2R	Normal
DoS	133070	10	0	0	2
Probe	20	1345	0	1	9
R2L	0	1	372	2	4
U2R	1	1	0	8	2
Normal	5	2	7	5	33100

TABLE IV CONFUSION MATRIX OF K2

	DoS	Probe	R2L	U2R	Normal
DoS	131568	534	7	798	207
Probe	2	1451	0	5	25
R2L	0	0	353	10	5
U2R	0	1	0	17	3
Normal	2	272	183	40	32484

Fig. 3 and Fig. 4 show the comparison number of correctly classified instances and incorrectly classified instances between K2 based IDS model and proposed REP based IDS model respectively.

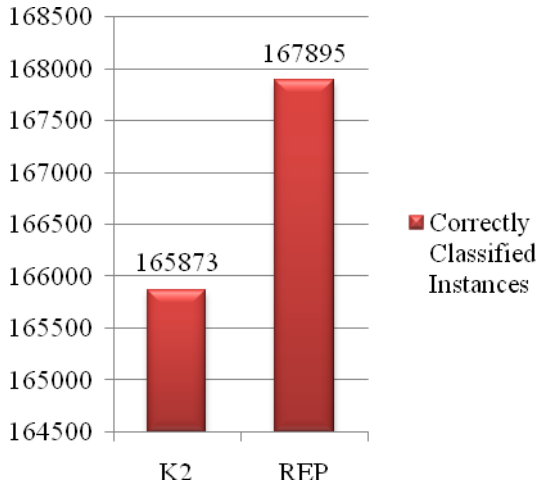

Figure 3: Correctly Classified Instances

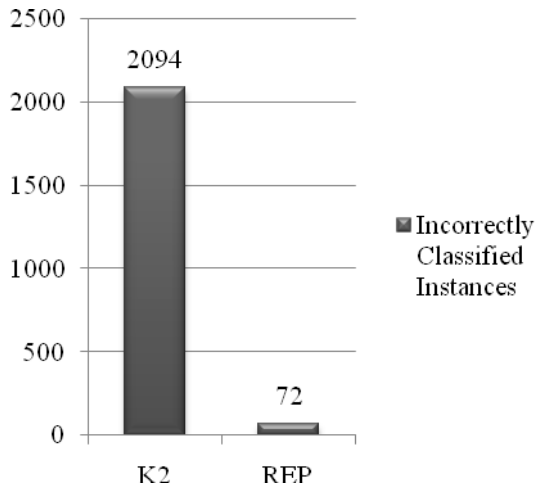

Figure 4: Incorrectly Classified Instances

The class wise comparison of accuracy between K2 based IDS model and proposed REP based model are shown in Fig.5 to Fig.9. It can be noticed than REP has been extremely efficient in detecting each type of attacks than K2 and only lacks in effectively detecting u2r attacks.

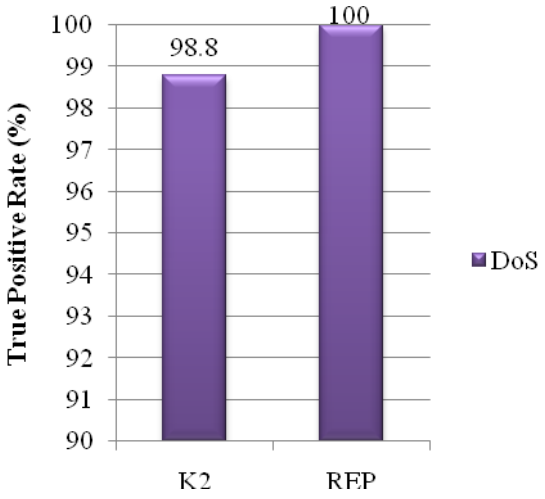

Figure 5: DoS Attacks Detected by K2 and REP

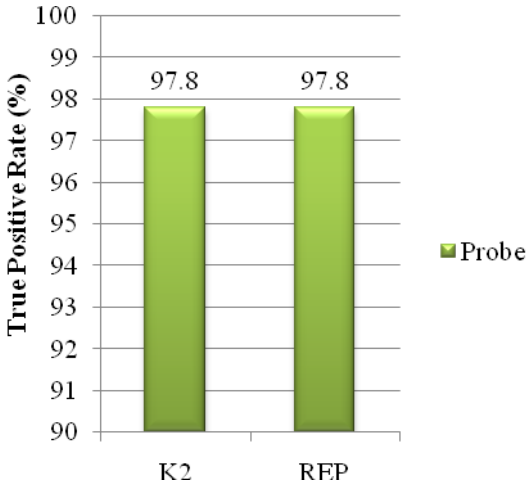

Figure 6: Probe Attacks Detected by K2 and REP

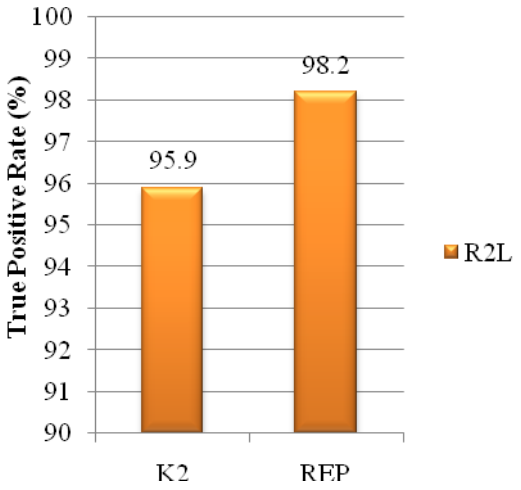

Figure 7: R2L Attacks Detected by K2 and REP

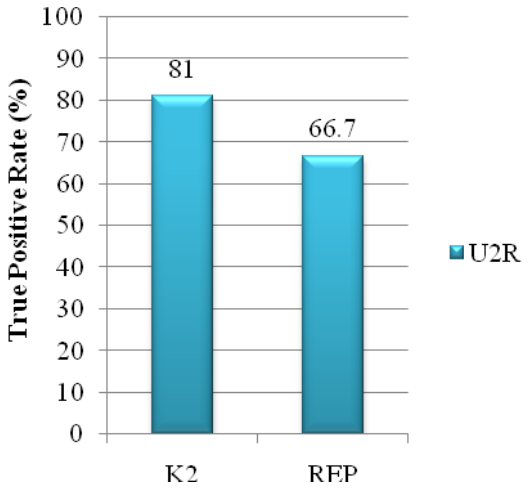

Figure 8: U2R Attacks Detected by K2 and REP

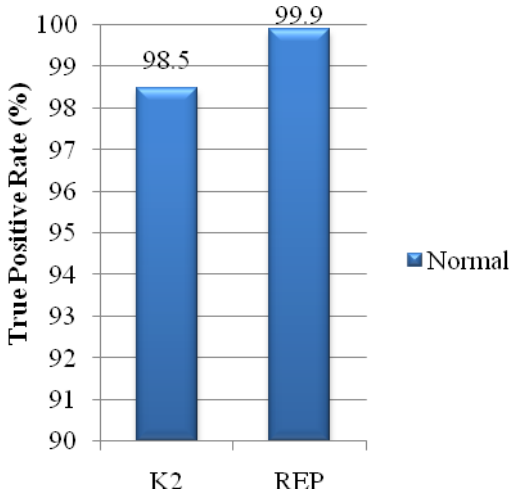

Figure 9: Normal Activities Detected by K2 and REP

The comparison of overall accuracy between K2 based IDS and REP based IDS models shows in Fig. 10. Again, REP excels K2 with higher accuracy.

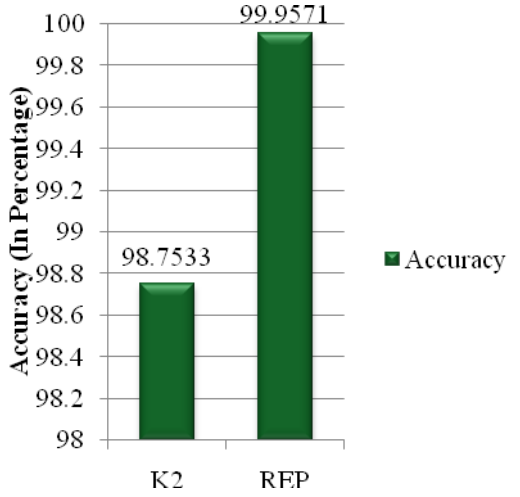

Figure 10: Accuracy of K2 based IDS and REP based IDS

When the proposed REP-based IDS model is compared against the commonly preferred techniques of Data Mining which were used for intrusion detection, the REP was found with the maximum accuracy result. The comparison can be visualized from Fig. 11:

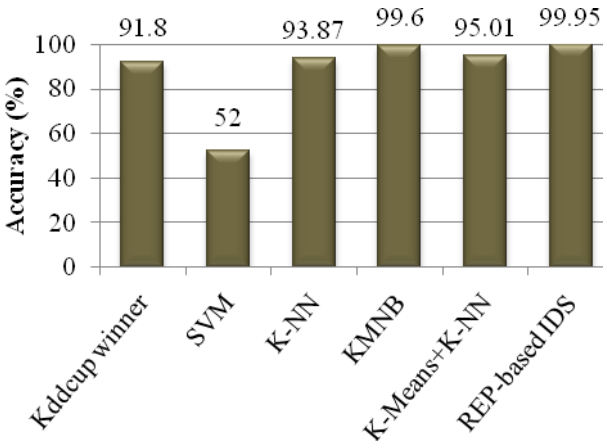

Figure 11: Accuracy of K2 based IDS and REP based IDS

Finally the Fig. 12 shows the comparison between the time consumed by the two algorithms to train and test the system. It can be observed that time consumed for training are satisfactory in both cases because it does not have much effect on the performance of the system while the test time in case of REP has come out as an advantage; as time required to test is crucial during evaluation.

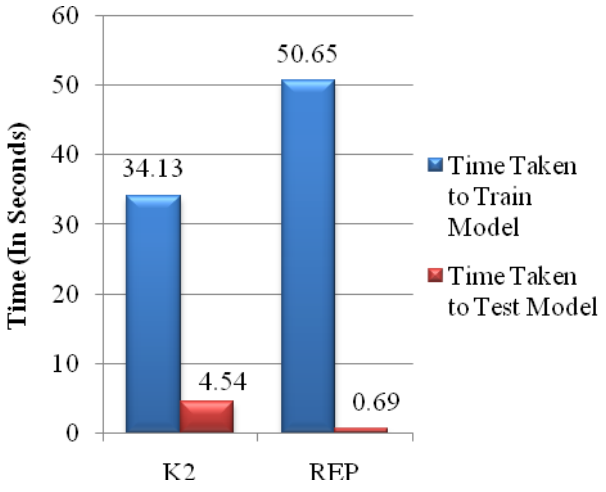

Figure 12: Time Consumed to Train and Test the Models

IV. Conclusion

The experimental analysis performed for intrusion detection has demonstrated the applicability of REP based Model excelling the compared frequently used K2 algorithm. Also the comparative tables and graphs shows, REP as the most efficient algorithm aiding in high accuracy and better detection in each type of attacks discussed in the paper. Also REP along with greater classification performance leads to reduce the error rate by

applying pruning on the dataset. Thus REP is supposed one of the preferable algorithms of the future. The only shortcoming faced or a future scope can be expected in reducing the higher amount of data for performing pruning.

REFERENCES

[1] Changxin Song, Ke Ma, "Design of Intrusion Detection System Based on Data Mining Algorithm," Proceedings of 2009 International Conference on Signal Processing Systems, IEEE 2009, pp. 307-373.

[2] Yogendra Kumar Jain and Upendra, "An Efficient Intrusion Detection based on Decision Tree Classifier Using Feature Reduction," International Journal of Scientific and Research Publication, Volume 2, Issue 1, pp. 1-6, January 2012.

[3] Manikandan R, Oviya P and Hemalatha C, "A New Data Mining Based Network Intrusion Detection Model," Journal of Computer Application, Volume 5, Issue EICA2012-1, pp. 1-10 February 10, 2012.

[4] Daejoon Joo, Taeho Hong and Ingoo Han, "The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors," Expert System with Applications 25, 2003, pp.69-75.

[5] Wenke Lee and Salvatore J.Stolfo, "Data Mining Approaches for Intrusion Detection," Proceedings of the 7th USENIX Security Symposium San Antonio, Texas, January 26-29, 1998.

[6] Evelina Lamma, Fabrizio Riguzzi and Sergio Storari, "Improving the K2 Algorithm Using Association Rule Parameters," Elsevier Publications, 2005, pp. 1-11.

[7] Tapio Elomaa and Matti Kääriäinen, "An analysis of Reduced Error Pruning," Journal of Atificial Intelligence Research 15, 2001, pp. 163-187.

[8] Johannes Fürnkranz, "Pruning Algorithms for Rule Learning," Machine Learning 27 Kluwer Academic Publishers, 1997, pp. 139-172.

[9] KDD Cup 1999 Data, Information and Computer Science, University of California, Irvine. http://kdd.ics.uci.edddatabases/kddcup99/kddcup99.html.

[10] Hesham Altwaijry and Saeed Algarny, "Bayesian based intrusion detection system," Journal of King Saud University – Computer and Information Sciences 24, 2012, pp. 1-6.

AUTHORS PROFILE

Mradul Dhakar is an M. Tech. Scholar, pursuing his post graduation in Computer Science and Engineering from Madhav Institute of Technology and Science, Gwalior, M.P. (India). He completed his graduation in 2011 in Computer Science and Engineering from Institute of Information Technology and Management, Gwalior, M.P. (India). His research interest includes Network Security and Data Mining.

Dr. Akhilesh Tiwari has received Ph.D. degree in Information Technology from Rajiv Gandhi Technological University, Bhopal, India. He is currently working as Associate Professor in the department of CSE & IT, Madhav Institute of Technology & Science, Gwalior, M.P. (India). His area of current research includes knowledge discovery in databases & data mining, and wireless Networks. He is also acting as a reviewer & member in editorial board of various international journals. He is having the memberships of various Academic/ Scientific societies including IETE, CSI, GAMS, IACSIT, and IAENG.

