ISSN:0975-9646

Intrusion Detection System using Ripple Down Rule learner and Genetic Algorithm

D.P. Gaikwad
Department of Computer Engineering
AISSMS College of Engineering
Pune, India

R.C.Thool
Department of Information Technology
SGGSI of Engineering and Technology
Nanded, India

Abstract—Intrusion detection system is used to identify anomalous packets in network. It can also identify unauthorized, malicious activity and malicious code in network. Currently, differnet apporaches of network intrusion detection systems are proposed by researchers. The classification based techniques has some issues such as model overfitting and classification evaluation. The challenging task in intrusion detection is to reduce the false positives and increase classification accuracy. The rule based techniques are simple, advanced and help to reduce the false positives. The rule-based intrusion detection systems and their performances mainly depend on the rule sets. But rules formation becomes a tedious and time consuming task due to the enormous amount of network traffic. In this paper, a novel architecture for intrusion detection system is presented which we call as RDRID. The RDRID is simple and advanved rule based intrusion detection system that reduce false positives and increase classification accuracy. In our implementation, we make use of Ripple Down Rule learner as classifier with Genetic Algorithm based features selection. The Genetic Algorithm is used to select the relevant features from training dataset. The performance of the proposed technique is evaluated in terms of classification accuracy, model building time and False Positive rates. The experimental results show that the proposed approach outperforms existing standard classifier.

***Keywords*—Ripple Down Rule, Genetic Algorithm, False Positive rate, Accuracy, Classification**

I. INTRODUCTION

The Internet is extensively growing in financial organizations, Universities, Information Technology industries and Government agencies. The electronic attacks on network and information system of the financial organizations, Military and Energy sectors are increasing day by day. Large numbers of computers in network becomes the victim of intruders. Intruders can have many forms such as viruses, spyware, worms, malicious logins and spamware. The secret information of any organizations may be leaked or damaged by intruder or unauthorized users. Intrusion detection system is used to spot and prevent the entrance of intruder in organization. It is used to identify the malicious use of computer and computer network. It detects access of unauthorized user, illegal users and the violation of security measures. There are two types of intrusion detection techniques viz., anomaly and misuse detection. Misuse detection is knowledge or pattern based, whereas anomaly detection is behavior based. Misuse detection is reliable for detecting known attacks with low false-positive, but it cannot detect the novel attack. Anomaly detection technique can detect novel attack with high false positive rate. The existing approaches of intrusion detection system have high detection rate, whereas they suffer from high false-alarms. The task of reducing warning is extremely necessary for intrusion detection system. The rule based intrusion detection system can help to reduce the false positive rates and increase classification accuracy. The objective of this research work is to reduce false positives and increase classification accuracy. We are able to achieve good classification accuracy using RDRID.

In this paper, we make two key contributions. Firstly, we have used Genetic Search Algorithm search to select the relevant features from NSL_KDD dataset. Then, we introduce a novel approach for anomaly based intrusion detection. The intrusion detection system is based on rule based classification technique that combines a Ripple Down Rule Learner and Genetic algorithm, which we call RDRID. The NSL_KDD dataset is used to generate rules using Ripple Down Rule learner. The proposed RDRID is evaluated by using Cross Validation of 10-fold and supplied test dataset. The performance of RDRID is measured in terms of classification accuracy, model building time and false positives. The experimental results are thoroughly described in this paper. The experimental results show that RDRID offers the best detection rate and classification accuracy. The rest of this paper is organized as follows. Section II provides an overview of related work. Section III describes the experiment methodology of the system. Section IV introduces the Ripple Down Rule Learner. Section V describes proposed architecture of intrusion detection system. In Section VI experimental Results are discussed. Finally, Section VII concludes the paper.

II. RELATED WORK

Heba Ezzat Ibrahim, Sherif M. Badr and Mohamed A. Shaheen [1] has proposed a multi-layer intrusion detection model using Naive Bayes, Multilayer Perceptron Neural Network and C4.5 decision tree Machine learning techniques. The experimental results show that classification rate accuracy of C4.5 decision tree is very high as compare to MPL and Naïve Bayes. The proposed model detects attack in first stage and it classifies attack in second stage. Sebastian Zander, Thuy Nguyen and Grenville Armitage [2] have proposed an unsupervised Machine learning technique

to classify traffic and identify application in network. Authors have used a feature selection technique to find out the optimal set of flow attributes. This statistical property of flow is used for classification and identification of packet in network. The influence of different attributes on the learning is also determined. Juvonen and T. Sipola [3] have combined unsupervised data analysis with rule extraction algorithm to implement online anomaly detection system. Conjunctive rule extraction algorithm is used to create rule sets. These rule sets are used to separate the training dataset into clusters. The same rule sets are used to classify traffic and detect intrusion in real time. Authors have also suggested that the combination of rule extraction algorithm and Machine learning methods is feasible solution to implement real time intrusion detection. Igino Corona, Giorgio Giacinto and Fabio Roli [4] have described a general taxonomy of attack tactics against intrusion detection system. The details about measurement, classification and response phases of IDS are given in this paper. The general discussion is found about Probably Approximately Correct algorithm, boosting and framework of robust statistics. The influence of single sample on the built model can be smoothed by using boosting. Authors have suggested that we can substitute statistical estimators such as the mean and variance by their robust versions to develop new learning algorithms. Arman Tajbakhsh, Mohammad Rahmati and Abdolreza Mirzaei [5] have used Association Based Classification for designing the intrusion detection system. The speed of Apriori algorithm is increased by reducing items involved in rule induction without any information loss. The fuzzy association rules are used to build descriptive models of different classes. The proposed classifier is efficient for classification of large dataset and can handles the symbolic attributes. Kamran Shaf and Hussein A. Abbass [6] presented UCSSE framework for real time extraction of maximum general rules using UCS (Supervised Learning Classifier) System supervised classifier. The algorithm used in this frame work automatically identifies and extract signatures for normal and intrusion activities. Muamer N. Mohammada, Norrozila Sulaimana and Osama Abdulkarim Muhsinb [7] have implemented intrusion detection system using data mining and expert system in WEKA. The performance, detection efficiency and false alarm rate are better than the existing system. Jiankun Hu and Xinghuo Yu [8] have implemented an enhanced incremental HMM stochastic process for intrusion detection system. The data preprocessing approach is used to speed up a hidden Markov model. The system is system-call based anomaly intrusion detection system. Mrutyunjaya Pandaa, Ajith Abrahamb and Manas Ranjan Patrac [9] have implemented the hybrid intelligent approach for intrusion detection system. The supervised or un-supervised classifier is used for data filtering. The output of this classifier is applied to another classifier to classify the training dataset. The system is very intelligent in decision and overall performance is enhanced. Mirco Marchetti, Michele Colajanni and Fabio Manganiello [10] have proposed framework which is based on two unsupervised classifier. Self-Organizing maps and pseudo-Bayesian, probability correlation algorithms are used to identify the multistep attacks. Shrinivasu and P.S.Avadhani [11] have proposed GA-NN based the intrusion detection system. Genetic Algorithm Weight Extraction Algorithm is used to extract and optimize the weights between the neurons of ANN to identify the intrusions effectively.

III. METHODOLOGY

The experimental setup is divided into two phases to train and evaluate the performance of proposed method of intrusion detection system. In the first phase, the data preprocessing technique is used to select relevant features. The dataset DARPA 1998, NSL-KDD99 and KDD99 can be used for generating rules for intrusion detection. In this paper, NSL-KDD99 dataset is used to generate rules for normal and abnormal packets in network. The preprocessing of dataset is essential to select relevant features is called as feature selection. In feature selection, subsets of relevant features are selected and irrelevant features are eliminated. Feature selection is essential to reduce dimension, boost generalization capability, accelerate learning and enhance model interpretation [15]. More feature selection may produce the problem of lack of generalization, whereas less feature selection causes degradation in level of classification quality. Selection of appropriate features in rules enables the rules to be more general [16][17]. Genetic Algorithm is applied on NSL_KDD99 dataset to select relevant features to enable rules more general. The Eighteen features out of Forty One features are selected by Genetic Algorithm.

In the second phase of the experiments, WEKA Data Mining tool was used to evaluate the performance of proposed rule based Classifier with other standard classifier. The performance of the proposed rule based classifier is evaluated by using the 10-fold Cross Validation and testing dataset. The performance of the intrusion detection approach is generally depends on the ability of the approach to distinguish normal vs. abnormal, time required for training the model and the time taken during the detection process. The proposed method is evaluated in term of model building time, False Positive rates and Classification accuracy. All experiments have performed on Lenovo Laptop with Intel(R) CORE™ i5-3210M CPU @ 2.50GHz, Installed 8GB RAM and 32 bit Operating system.

IV. INTRODUCTION TO RIPPLE DOWN RULE LEARNER

Ripple Down Rule is a knowledge acquisition and representation methodology proposed by Compton and Jansen. The knowledge is rapidly acquired and maintained by the domain expert using Ripple down rule. The knowledge is acquired based on the current context and is added incrementally. It is an interesting representation scheme for knowledge acquisition. It consists of a data structure and knowledge acquisition scenarios. The data structure is used to store Human expert's knowledge in coded form is called as rule. A ripple down rule is a list of rules, each of which may be connected to another ripple

down rule, specifying exceptions. An expert can create a new rule and insert in list, which only have effects in the given context of the parent rule. The ripple down rule technique creates a two-way dependency relation between rules such that rule activation is investigated only in the context of other rule activation. Thus, ripple down rules form a binary decision tree that differs from standard decision trees in which all decisions are made at root nodes [12]. Default rule are generated first and then the exceptions for the default rule with the least error rate. It performs a tree-like expansion of exceptions. A set of rules that predict classes other than the default is called as exception rules. The rules are never modified or deleted but they are locally patched. The approach is based on the combined use of cases to motivate knowledge acquisition and validation. The knowledge is acquired as production rules and structured into decision lists of exceptions. A key feature of RDR, and the reason why maintenance is easily managed, is that rules are never modified or deleted but they are locally patched [13]. That is, new rules are exceptions to previous rules and the new rule is validated in the context of previously seen cases [14]. Rules and their exceptions are ordered, and the first condition that is satisfied fires the corresponding rule.

V. PROPOSED ARCHITECTURE OF INTRUSION DETECTION SYSTEM

In this section, a novel architecture of RDRID intrusion detection is discussed in detail. The proposed architecture (shown in Fig.1) is composed of two phases: one is training phase and other is detection phase. In first phase, the Genetic Algorithm is applied on NSL_KDD dataset to select relevant features to improve the classification accuracy. Then, Ripple Down Rule Learner is used to mine the rules for intrusion detection system. The second phase of architecture is used to detect the on line packets. Firstly, the detection phase is used to capture the on line packets over network using packet sniffer and then preprocessed. The preprocessed packets are sent to model for detection. The packets are matched with the rules of model and detected as normal, abnormal according to the rule matching. Following 2 subsections describes the working principle of proposed RDRID intrusion detection system in detail.

A. Feature Selection using Genetic Algorithm

The relevant feature selection is a main area of research in intrusion detection system. Preprocessing of the dataset can be used to select relevant features and reduce the dimensionality of dataset. The most of the features indicated in NSL_KDD99 dataset are not necessary for classification of packets. If more number of features selected then more model building time is required. The relevant feature selection is the most challenging task in intrusion detection system. The rule based classifier is mainly concerned with the feature selection problem. It is concerned with identifying the features of the data that experts wish to use in rules. The NSL-KDD99 dataset have suggested

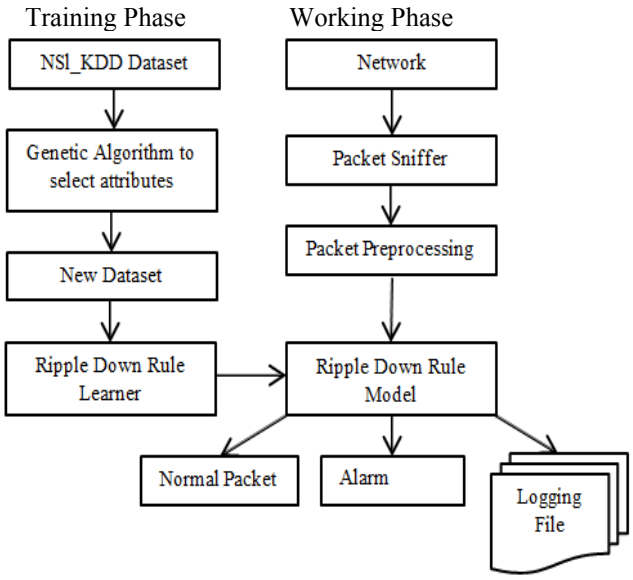

Fig 1.Proposed System Architecture

Forty One feature to implement intrusion detection system. All Forty One features are not essential for packet classification. In this paper, we make use of Genetic Search Algorithm to select relevant feature from NSL_KDD dataset to increase classification accuracy and reduce the model building time. Eighteen features have selected using Genetic Search Algorithm. The relevant features are listed in table 1. The Genetic Search Algorithm is applied with following parameters.

- Population size : 20
- Number of generations : 20
- Probability of crossover : 0.7
- Probability of mutation : 0.033
- Report frequency : 20
- Random number seed : 01

TABLE 1. SELECTED FEATURES BY GENETIC ALGORITHM

Sr.No.	Selected Features Using Genetic Algorithm
1	Service
2	Flag
3	src_bytes
4	dst_bytes
5	wrong_fragment
6	logged_in
7	su_attempted
8	is_host_login
9	srv_serror_rate
10	rerror_rate
11	same_srv_rate
12	diff_srv_rate
13	dst_host_count
14	dst_host_same_srv_rate
15	dst_host_diff_srv_rate
16	dst_host_srv_diff_host_rate
17	dst_host_serror_rate
18	dst_host_srv_serror_rate

B. Geneation of Rules using Ripple Down Rule learner

In this paper, we make use of the Ripple Down Rule learner to generate set of rules for normal and abnormal packets. Rule learner generates Forty Seven rules including the default and exception rules. The sample rules generated by Ripple Down Rule learner are listed in table 2. These rules are used to build model for intrusion detection system.

TABLE 2. SAMPLE RULES GENERATED BY RIPPLE DOWN RULE LEARNER

Sr. No.	RIpple DOwn Rule Learner rules
1	Except (src_bytes > 28.5) and (src_bytes <= 333.5) and (dst_bytes > 28.5) and (service = http) and (src_bytes > 142.5) => class = normal
2	Except (src_bytes > 28.5) and (dst_bytes > 0.5) and (dst_bytes <= 2181) and (rerror_rate <= 0.005) and (dst_bytes <= 374.5) and (dst_host_same_srv_rate > 0.175) and (dst_host_count > 5.5) and (dst_bytes <= 206.5) => class = normal
3	Except (src_bytes > 28.5) and (dst_host_diff_srv_rate > 0.025) and (dst_bytes > 0.5) and (dst_bytes <= 374.5) and (src_bytes > 104) => class = normal
4	Except (src_bytes > 8.5) and (logged_in = 1) and (src_bytes <= 16869.5) and (dst_host_count > 16.5) and (dst_bytes <= 2184.5) and (dst_host_count <= 177.5) and (dst_host_count > 52.5) => class = normal
5	Except (src_bytes > 8.5) and (logged_in = 1) and (src_bytes <= 36148) and (dst_host_srv_diff_host_rate <= 0.065) and (dst_bytes <= 2178.5) and (src_bytes > 334.5) and (dst_bytes <= 488.5) and (dst_host_srv_diff_host_rate <= 0.015) => class = normal
6	Except (src_bytes > 8.5) and (src_bytes <= 519.5) and (src_bytes > 28.5) and (dst_host_count > 16.5) and (dst_bytes > 234) and (dst_host_same_srv_rate > 0.99) and (src_bytes > 141.5) => class = normal
7	Except (same_srv_rate > 0.495) and (src_bytes > 28.5) and (src_bytes <= 519.5) and (src_bytes <= 206.5) and (dst_bytes <= 166) and (dst_host_count > 230.5) and (dst_host_same_srv_rate > 0.005) => class = normal
8	Except (dst_host_count <= 254.5) and (src_bytes > 28.5) and (dst_host_diff_srv_rate > 0.015) and (service = ftp_data) and (src_bytes > 353.5) and (dst_host_srv_diff_host_rate <= 0.095) => class = normal
9	Except (diff_srv_rate <= 0.015) and (service = http) and (dst_host_srv_diff_host_rate > 0.005) and (dst_host_srv_diff_host_rate <= 0.345) => class = normal
10	Except (src_bytes > 8.5) and (dst_host_same_srv_rate <= 0.305) and (wrong_fragment <= 0.5) and (src_bytes <= 979) and (src_bytes > 103.5) and (src_bytes <= 229.5) => class = normal

VI. EXPERIMENTAL RESULTS AND DISCUSS

Once we build rule model, it can be used to detect abnormal packets in network. The performance of RDRID for intrusion detection system is calculated on the basis of number of rules. The performance is measured according to the classification accuracy, model building time and false positive rate. The classification accuracy is measured using the following expressions.

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

Where, FN is False Negative, TN is True Negative, TP is True Positive, and FP is False Positive. The false positive rate is the number of normal packets that are misclassified as attacks divided by the number of normal packets. The RDRID is evaluated in term of classification accuracy, model building time and false positives. The performance of RDRID along with existing classifiers using Cross Validation of 10-fold are listed in table 3. Fig.2 shows comparision of RDRID with other classifier. According to Fig. 2, the classification accuracy of RDRID using Cross Validation of 10-fold is more as compare with other existing classifier.

TABLE 3.COMPARISION OF CLASSIFIER ACCURACY USING CROSS VALIDATION OF 10-FOLD

Name of Classifier	Time taken to build model (Sec.)	False Positives	Classifier Accuracy In %
Naïve Bays	275.67	0.102	90.2233
Random Tress	36.39	0.003	99.661
C4.5/J48	418.79	0.004	99.6475
AdaBoost(base Classifier DecisionStump)	804.29	0.065	94.0035
Bagging(Base Classifier REPTree)	41.59	0.003	99.6761
Proposed RDRID	2337.3	0.003	99.719

The performances of RDRID along with existing classifiers using supplied test dataset are listed in table 4. Fig. 3 shows comparision of RDRID with other classifiers. According to Fig.3, the classification accuracy of RDRID using supplied test dataset also is more as compare with other existing classifier.

TABLE 4.COMPARISION OF CLASSIFIER ACCURACY ON TSET DATASET

Name of Classifier	Time taken to build model (Sec.)	False Positives	Classifier Accuracy In %
Naïve Bays	277.92	0.194	76.6812
Random Tress	36.47	0.193	77.1735
C4.5/J48	440.6	0.189	76.2287
AdaBoost(base Classifier DecisionStump)	835.61	0.199	74.583
Bagging(Base Classifier Random Tree)	300.32	0.159	79.8749
Propose RDRID	2337.3	0.153	80.7709

The confusion matrix is used a measurement criteria of proposed rule based classifier. The confusion matrix of RDRID on testing data is given in table 5.

TABLE 5. CONFUSION MATRIX OF PROPOSED RULE LEARNER

Classified As	Normal	Anomaly
Normal	67254	89
Abnormal	265	58365

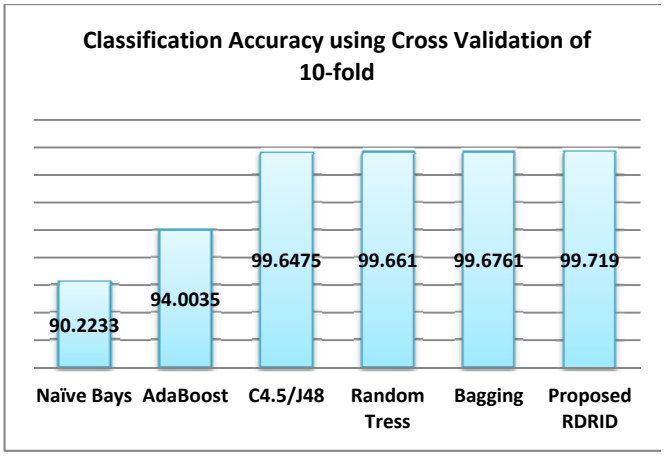

Fig.2 Classification Accuracy Cross Validation

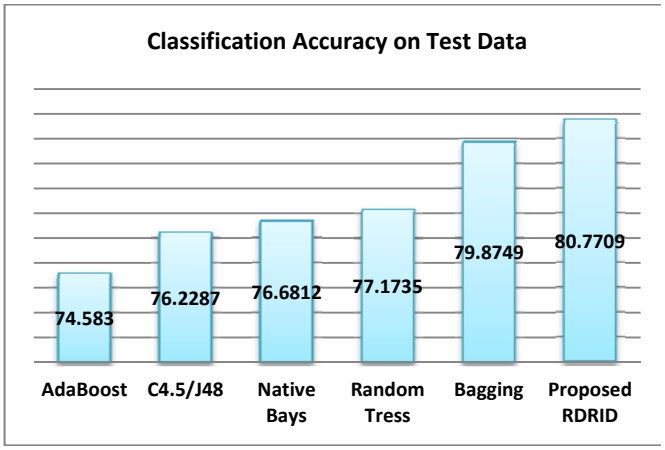

Fig.3 Classification Accuracy using testing dataset

VII. CONCLUSIONS

In this paper, we present the Ripple Down Rule learner with Genetic Algorithm for intrusion detection system to improve classification accuracy and reduce false positive rate. To increase the rate of accuracy, the Genetic Algorithm is used for feature selection. The selected features by Genetic Algorithm are presented in the paper. These features are used to generate the rule set for anomaly and normal packets in network. Ripple Down Rule learner is used to generate the set of rules. Experiments are performed using RDRID learner. All experimental results in this paper are based on our proposed feature selection method using Genetic Algorithm. In addition, we have compared the classification accuracy of RDRID against four exiting approaches. The experimental results show that proposed approach for intrusion detection outperforms them. We have achieved an average classification accuracy of 80.7709 %, and False Positive rate of 0.153 on testing data. Further, Cross Validation evaluation method show that classification accuracy of proposed approach is near 99.719 % and false positive rate of 0.003. This detection rate is possible due to the eighteen data attributes selected by Genetic Algorithm. The proposed classifier exhibits more model building time as compare to existing classifier. In addition, future work is concerned with decreasing the model building time of the proposed system.

REFERENCES

[1] Heba Ezzat Ibrahim, Sherif M. Badr and Mohamed A. Shaheen, "Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems", in International Journal of Computer Applications Vol. 56, No.7, October 2012.

[2] Sebastian Zander, Thuy Nguyen and Grenville Armitage, "Automated Traffic Classification and Application Identification using Machine Learning", in Proceedings of the IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05) , 2005.

[3] Juvonen and T. Sipola,"Combining conjunctive rule extraction with diffusion maps for network intrusion detection", in the Eighteenth IEEE Symposium on Computers and Communications (ISCC 2013), IEEE 2013.

[4] Igino Corona, Giorgio Giacinto and Fabio Roli, "Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues", in Inform. Sci. (2013).

[5] Arman Tajbakhsh, Mohammad Rahmati and Abdolreza Mirzaei," Intrusion detection using fuzzy association rules," in journal Applied Soft Computing 9 (2009) 462–469, Published by Elsevier.

[6] Kamran Shaf and Hussein A.Abbass, "An adaptive genetic-based signature learning system for intrusion detection," in Expert Systems with Applications 36, 2009.

[7] Muamer N. Mohammada, Norrozila Sulaimana and Osama Abdulkarim Muhsinb, "A Novel Intrusion Detection System by using Intelligent DataMining in WEKA Environment:," in Procedia Computer Science 3 (2011) 1237–1242. 2011.

[8] Jiankun Hu and Xinghuo Yu, "A Simple and Efficient Hidden Markov Model Scheme for Host-Based Anomaly Intrusion Detection", in IEEE Network January/February 2009.

[9] Mrutyunjaya Pandaa, Ajith Abrahamb and Manas Ranjan Patrac, "A Hybrid Intelligent Approach for Network Intrusion Detection", in International Conference on Communication Technology and System Design 2011.

[10] Mirco Marchetti, Michele Colajanni and Fabio Manganiello,"Framework and Models for Multistep Attack Detection", in International Journal of Security and Its Applications Vol. 5 No. 4, October, 2011.

[11] P. Shrinivasu and P.S.Avadhani, "Genetic Algorithm based Weight Extraction Algorithm for Artificial Neural Network Classifier in intrusion Detection", in Procedia Engineering 38 (2012) 144 – 153, Published by Elsevier Ltd. (2012).

[12] Priyanka Sharma, "Ripple-Down Rules for Knowledge Acquisition in Intelligent System", (JTES) Delving: Journal of Technology and Engineering Sciences Vol. 1, No. 1 January –June 2009.

[13] Compton, P., Preston, P. and Kang, B., (1994) Local Patching Produces Compact Knowledge Bases A Future in Knowledge Acquisition (eds) L. Steels, G. Schreiber and W. Van de Velde, Berlin, Springer Verlag, 104-117.

[14] Hendra Suryanto, Debbie Richards and Paul Compton, "The Automatic Compression of Multiple Classification Ripple Down Rule", Macquarie University, Sydney, Australia.

[15] G.Prashanth, Prashanth, P.Jayashree and N.Srinivasan, "Using Random Forests for Network-based Anomaly detection at Active routers", IEEE-International Conference on Signal processing, Communications and Networking Madras Institute of Technology, Anna University Chennai India, Jan 4-6, 2008. Pp93-96

[16] P.Compton, G.Edwards, B.Kang, L.Lazarus, R. Malor, T.Menzies, P.Preston. A.Srinivasan and S.Sammut, "Ripple down rules: possibilities and limitations", School of Computer Science and Engineering, University of New South Wales,PO Box 1, Kensington NSW, Australia 2033.

[17] Matjaž Juršič, Igor Mozetič and Nada Lavrač, "Learning Ripple Down Rules for Efficient Lemmatization", Department of Knowledge Technologies, Jožef Stefan Institute Jamova 39, 1000 Ljubljana, Slovenia.