Performance Analysis of Pre-Processing Techniques with Ensemble of 5 Classifiers

Sonali Kadam
Department of Computer Engineering
Bharati Vidyapeeth's college of Engg for Women.
Pune,India
kadamsonali@rediff.com

Rutuja Pawar
Department of Computer Engineering
Bharati Vidyapeeth's college of Engg for Women
Pune,India
pawarrutuja845@gmail.com

Manisha Kumari
Department of Computer Engineering
Bharati Vidyapeeth's college of Engg for Women.
Pune,India
mannu.kri1234@gmail.com

Shweta Phule
Department of Computer Engineering
Bharati Vidyapeeth's college of Engg for Women.
Pune,India
shweta1995phule@gmail.com

Priyansha Kher
Department of Computer Engineering
Bharati Vidyapeeth's college of Engg for Women.
Pune,India
priya98kher@gmail.com

Abstract—The continuous development in network attack is being a difficult issue in programming industry. Intrusion detection framework is utilized to identify and break down system attack so IDS need to be updated that can screen the framework and can trigger the alert in the framework. Numerous methods have been proposed by various authors to enhance the execution of IDS yet at the same time they can't give legitimate or complete solution.In the proposed work authorsconsidered several classification techniques and selected the most suitable classifiers namely Bayesian Network, Naive bayes, JRip, MLP, IBK, PART and J48 based on the accuracy.These selected classifiers were further ensemble and experiments were performed on the combination of ensemble of classifiers. The combination giving best accuracy will be used in IDS for detection of various attacks. In additiontwo pre-processing techniques were used for the performance analysis. The outcome of these experiment shows improvement in the detection rate of U2R and R2L attack.

Keywords*-bayesian network, intrusion detection system, ibk, jrip, mlp, naïve bayes, part and j48;*

*****

INTRODUCTION

Intrusion Detection framework is utilized to shield PC framework from the risk of theft from intruders. In late year web security volume and refined target attack has been expanded generously. There is increment in number of dangers and vulnerabilities like-business system framework, military and so on, which drives Intrusion Detection System as a noteworthy research region. Intrusion Detection System is isolated into two sections –Anomaly-based and Misuse-based [1]. Anomaly-based model is utilized for the deviation of new information from the pre-characterized profile of information. Misuse based is also called knowledge based, is utilized to perform location by means of coordinating the new information with the known assaults in database. Numerous scientists have proposed different information digging systems for Misuse-based IDS.

Machine learning techniques are used to build a model for predictions [3]. Different techniques like single classifier, Hybrid approach, Multi-classifier approach, ensemble approach etc. can be used.Ensemble is one of the method where different classifiers are combined in order to achieve good results[1]. The primary point of this technique is to build the exactness by joining distinctive classifiers. To perform ensemble of classifiers different methods are available. It is seen that accuracy of classifiers increments when combined together. The drawback of ensemble is classifier consume more time for learning and require more memory for processing.

In the proposed idea authors studied various classifiers and its performance. On the basis of literature survey, author considered some of the best classifiers to ensemble. Further evaluation of individual classifiers were performed and the results were analysed. The classifiers giving the best detection rate individually were selected for ensemble. Hence, the classifiers ensured are Naïve Bayes, Bayesian Net, PART, IBK, MLP and J48.Previously this combination of classifiers were not considered for evaluation.

Many authors have performed experiments using single pre-processing techniques with different combinations of classifier.Selection of pre-processing technique is one of an important role in order to handle huge dataset. Hence to know the strength of each pre-processing technique authors will evaluate the performance of 2 pre-processing techniques-

Normalization and Discretization. The performance analysis of both the techniques will be performed by authors.

The rest of the paper is partitioned into four areas – Section 2 gives description of how ensemble methods were used, Section 3 depicts the classifiers in proposed framework. In Section 4 proposed framework is depicted. In Section 5 conclusion and future extension is mentioned.

II. BACKGROUND

In recent years, an ensemble strategy is vastly used to recognize intrusion in the framework. In 2013,Himadri Chauhan, Vipin Kumar et al. determination of main 10 classifiers-SGD, IBK, JRip, PART, J48, Random Tree, Logistics, Bayes Net were finished. At first many experiment were performed on these grouping calculations and after that they were picked. Out of those J48, IBK gave better outcomes as for exactness and less handling time [6]. In 2015, Sumoli Choudhary and Anirban Bhouwan proposed a framework with a few arrangements strategies and machine learning calculations. The classification techniques used where Bayes Net, IBK, JRip, PART, Logistic, J48, Random Tree, J48 and REPtree. All these were gathered with machine calculations Boosting, Bagging and Stacking approach. Henceforth the outcomes demonstrated that Bayes Net and J48 Tree are best order strategies and boosting gives great outcome [2].In 2015 Kailas Elekar et. al. analysed rule based order systems – Decision Tree, PART, ZeroR, OneR, and JRip. The execution of all these arrangement strategies evaluated on the basis of cross validation and test data from which, PART gave better results than others [4]. In 2013 Vijay Katkar and Siddhart Kulkarni proposed a framework for discovery of Denial of Service assault in system. The classifiers utilized for examinations were Naïve Bayesian, Bayesian Network, Sequential Minimal Optimization, J48, and REPtree. After performing analyses the outcomes expressed that outfit of J48, REPtree and Bayesian Network recognized DoS attack with critical precision. The authors likewise demonstrated that group of classifier can be utilized as opposed to building new classifier [1]. In 2015 author Ployphan Sornsuwit and Saichon Jaiyen proposed an Intrusion recognition based model for identification of U2R and R2L assaults. The authors embraced Adaboost calculation for gathering of feeble/weak learners and to help the execution. The powerless learners utilized were Decision Tree, SVM, Naïve Bayes and MLP. Furthermore, authors acknowledged the utilization of relationship based calculations for lessening of elements in dataset. The outcomes demonstrated that troupe of Naïve Bayes and MLP gave great outcomes for U2R and R2L attack with most noteworthy affectability. Choice Tree bombed for this situation [3]. In 2014 Tanya Garg and Surinder Singh Khurana performed correlation of various grouping strategies for Intrusion Detection System. Creators utilized Garette positioning strategy to rank the classifiers. As indicated by Garette positioning Rotation Forest was positioned.

III. THEORETICAL PRELIMINARIES

A.Pre-processing Techniques:

- *Normalization:*

Normalization is one of the pre-preparing method which scales up and downsizes the information i.e. control of information is done before it is utilized as a part of further stages. Normalization has numerous systems like Min-Max standardization, Z-score standardization, Integer Scaling Normalization and Decimal scaling standardization. In min-max relationship is kept up between unique information and it gives straight change. Fundamental point of utilizing this system is that it deciphers the outcome precisely, remove variations and exceptions, and reveal the patterns, and so on.

- *Discretization:*

Discretization is utilized as a pre-preparing venture for machine learning calculations that handle just discrete information. Discretization additionally goes about as an element determination strategy that can altogether affect the execution of characterization calculations which is utilized as a part of the examination of high-dimensional biomedical information. It has crucial implication for the investigation of high dimensional genomic and proteomic information. It is the way toward changing a continuous-valued variable into a discrete one by creating a set of contiguous intervals or equivalently a set of cut points that explain the range of the variable's values. Discretization techniques fall into two classes: unsupervised, which don't utilize any data in the objective variable and administered strategies, which do. It has been measured that directed discretization is more advantageous to arrangement than unsupervised discretization. Administered discretization strategies will discretize a variable to a solitary interval if the variable has almost no relationship with the objective variable. This permanently removes the variable as input to the classification algorithm.

B. Classifiers:

- *Naïve Bayes:*

Naive bayes is a basic system which assigns out class labels to issue cases. Naive Bayes is one of the classifier in which each element has its own particular free esteem which is not subject to other element. Naive bayes classifier can get rid of vast scale arrangement issues regardless of the possibility that the whole preparing set is not fitted in the memory. Points of interest of naive bayes is that it is best for spam filtering and archive arrangement.

- *J48:*

J48 classifier is otherwise called C4.5 choice tree which is utilized for arrangement. Utilizing this classifier, binary tree can be built. Missing values are overlooked by J48 i.e. it predicts the qualities for these traits in light of the preparation dataset. The fundamental thought is that it isolates the information into ranges. Classification is finished utilizing choice tree or its rules created in J48.

- *IBK:*

Fundamentally "IB" remains for Instance-Based and "k" determines number of neighbors that are analysed. IBK executes k-Nearest Neighbor calculation. In IBK, information is spoken to in a vector space. It is utilized for characterization, relapse and evaluating constant factors. In light of cross approval it can choose suitable estimation of K. Distance weighting can also be done.

- *JRip:*

It is one of the fundamental and most well-known calculations. Every one of the classes are analysed as the developing size and the underlying arrangement of rules are produced by utilizing the decreased error rate. In JRip every one of the cases continues by a specific choice in preparing information as a class, and it finds a decision that cover every one of the individuals from a class. After that it continues, to the following class and this strategy goes ahead till the end, until all classes are secured properly.

- *PART:*

It is known as the separate and-Conquer administer learner. It delivers all the arrangement of principles called as 'Choice list'. As the principles are created, another arrangement of information is contrasted with each administer and after that the things are allocated to the class of first coordinating standard. It manufactures a halfway c4.5 choice tree in each emphasis and makes the best leaf into a rule.

- *MLP:*

Multi-Layer Perceptron (MLP) is a system of simple neurons which is called as perceptron. MLP is a feed forward neural system with at least one layers amongst information and yield layer. Sustain forward implies that information streams in one direction from input to output layer (forward bearing). This sort of system is prepared with the back spread learning calculation. MLPs are generally utilized for example acknowledgment, characterization, forecast and guess. The issues which are not directly distinct can be settled by Multi-Layer Perceptron. The perceptron figures a single output from various genuine valued information sources. It figures yield by shaping a straight mix as indicated by its input weights. And afterward putting the yield through some nonlinear activation function. The quantity of hidden units can likewise be determined.

- *Bayes Network:*

Bayes net (Bayesian Network) is a classifier in based on probabilistic model. It speaks to an arrangement of arbitrary factors with their restrictive conditions through a coordinated non-cyclic chart or directed acyclic graph (DAG). It utilizes different pursuit calculations and quality measures. Bayes net just relates that specific hub which is probabilistically related by some easy-going conditions which gives an immense sparing of calculation. There is no compelling reason to store all the conceivable arrangement of express, the main thing has expected to relate with all conceivable related blend of sets. This makes an immense sparing of calculation and space table. Another motivation to utilize Bayes net is that they are versatile. It begins with a little and restricted information about a space and procure new learning. Along these lines, one's need not need to keep finish learning about the occurrence or space. Preferred standpoint of utilizing Bayes net is that probabilities require not being correct. It approximate to surmised probabilities. It utilizes easy-going restrictive probabilities than switch to assess. Since it is ideal to estimate probabilities "in the forward direction".

IV. PROPOSED SYSTEM

According to the literature survey, classifiers are being selected for detecting various attacks. Classifiers having good accuracy are selected for ensemble so that better results are obtained.The main advantage of proposed idea is, implementation of such classifiers will be done which can detect all the attacks. Depending on the survey we can conclude that different classifiers are responsible for detection of different attacks.

Adoption of seven different classifiers such as Bayesian Network, Naive bayes, PART, JRip, MLP, J48 and IBK is being done in the experiment.In proposed system, experiments will be performed on total 21 combinations of classifier. Each combination will be executed with respect to both the preprocessing techniques. Hence overall 42 experiments will be performed. Once all the combinations are executed, results are analyzed with respect to accuracy, correctly classified attacks and class wise attack detection.

Prior to ensemble pre-processing techniques are implemented. Due to pre-processing accuracy increases after datacleaning. Hence for increasing accuracy and results in the proposed method two techniques were used -Normalization and Discretization. Pre-processing techniques are performed on both training and test dataset. Hence in proposed system, ensemble of such classifiers is done with pre-processing techniques for acquiring best results.

Aim of the proposed system is to detect attacks like Normal, Probe, U2R, R2L, DoS. In earlier systems detection of DoS attack is done 100% hence detection of U2R and R2L attack will be increased in this proposed system. Subsequently identifying all assaults in system will help in prevention of various illegal practices. Thus it will be helpful in Intrusion Detection Systems. As the greater part of the security can be kept up by IDS it's a need in today's world to have good IDS.

Working of system is shown in Figure 1.Initially the dataset is divided into 2 parts: Training dataset and Testing dataset. Training dataset is used to train the classifiers for prediction. Once the classifier is trained test dataset is given as an input to classifiers and further the classifiers predict the attacks. The train and test dataset is pre-processed .The dataset is further given to the 5 classifiers .Once all the classifiers predicts its output voting is performed on the output and is given as final predicted class. All this process will be executed on 21 combinations of classifiers. The best combination will be selected for further implementation of IDS.

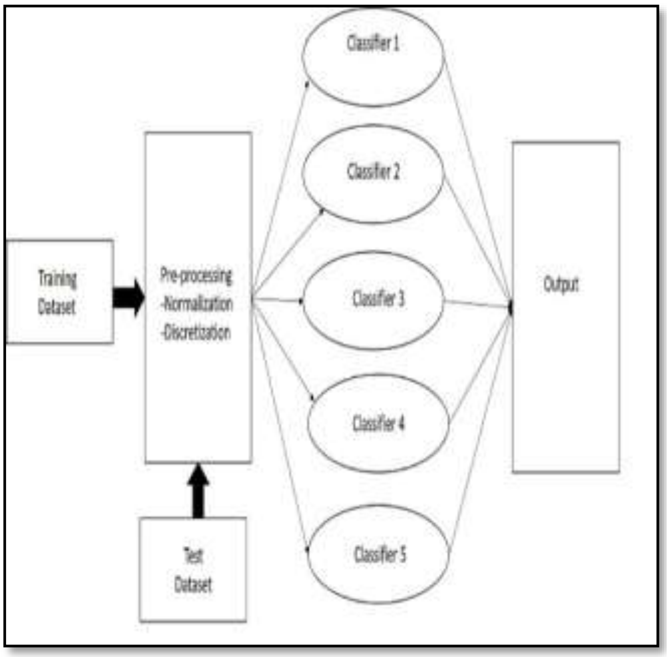
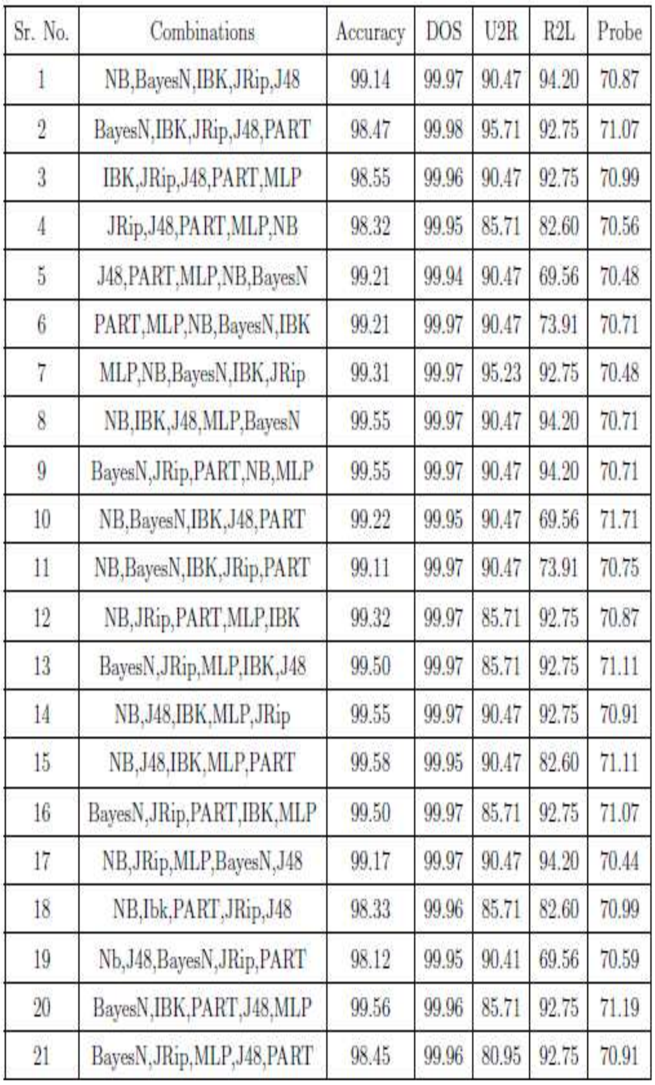


Table 1:Results of Normalization technique.



Sr. No.	Combinations	Accuracy	DOS	U2R	R2L	Probe
1	NB,BayesN,IBK,JRip,J48	99.14	99.97	90.47	94.20	70.87
2	BayesN,IBK,JRip,J48,PART	98.47	99.98	95.71	92.75	71.07
3	IBK,JRip,J48,PART,MLP	98.55	99.96	90.47	92.75	70.99
4	JRip,J48,PART,MLP,NB	98.32	99.95	85.71	82.60	70.56
5	J48,PART,MLP,NB,BayesN	99.21	99.94	90.47	69.56	70.48
6	PART,MLP,NB,BayesN,IBK	99.21	99.97	90.47	73.91	70.71
7	MLP,NB,BayesN,IBK,JRip	99.31	99.97	95.23	92.75	70.48
8	NB,IBK,J48,MLP,BayesN	99.55	99.97	90.47	94.20	70.71
9	BayesN,JRip,PART,NB,MLP	99.55	99.97	90.47	94.20	70.71
10	NB,BayesN,IBK,J48,PART	99.22	99.95	90.47	69.56	71.71
11	NB,BayesN,IBK,JRip,PART	99.11	99.97	90.47	73.91	70.75
12	NB,JRip,PART,MLP,IBK	99.32	99.97	85.71	92.75	70.87
13	BayesN,JRip,MLP,IBK,J48	99.50	99.97	85.71	92.75	71.11
14	NB,J48,IBK,MLP,JRip	99.55	99.97	90.47	92.75	70.91
15	NB,J48,IBK,MLP,PART	99.58	99.95	90.47	82.60	71.11
16	BayesN,JRip,PART,IBK,MLP	99.50	99.97	85.71	92.75	71.07
17	NB,JRip,MLP,BayesN,J48	99.17	99.97	90.47	94.20	70.44
18	NB,Ibk,PART,JRip,J48	98.33	99.96	85.71	82.60	70.99
19	Nb,J48,BayesN,JRip,PART	98.12	99.95	90.41	69.56	70.59
20	BayesN,IBK,PART,J48,MLP	99.56	99.96	85.71	92.75	71.19
21	BayesN,JRip,MLP,J48,PART	98.45	99.96	80.95	92.75	70.91

Figure1.SystemArchitecture.

V.EXPERIMENTS AND RESULTS

In the proposed idea, KDD'99 Cup dataset is used for analysis and the compatible classification algorithms have been analyzed using WEKA tool. 2 pre-processing techniques-Normalization and Discretization are used to handle large dataset and to increase the performance. Consequently combination of classifier giving great accuracy is grouped in proposed framework. Hence, ensemble of Bayesian classifier, Naïve bayes, IBk, JRip, MLP, J48 and PART will be performed. Classifier will be ensemble using Voting technique to predict the accuracy based on voting factor.

Table 1 shows all the results of 21 combination with Normalization as the preprocessing technique. From the table it is clear that the accuracy gained is best. Even the detection rate of DoS attack is best by all the combinations.
Different combinations 1, 8, 9, 17 gave best detection rate for R2L i.e.:94.20.Similarly combinations 5, 6, 8,9,10,11,14,15 gave highest detection rate for U2R attack from Table 1.
Hence overall the highest accuracy was gained by combination Naive Bayes, IBK, J48, Multilayer Perceptron,Bayes Net i.e 99.5.By the same combination high detection of R2L attack was achieved i.e.94.20 from Figure 2.

The results obtained using both the preprocessing techniques and 21 combinations are shown below.

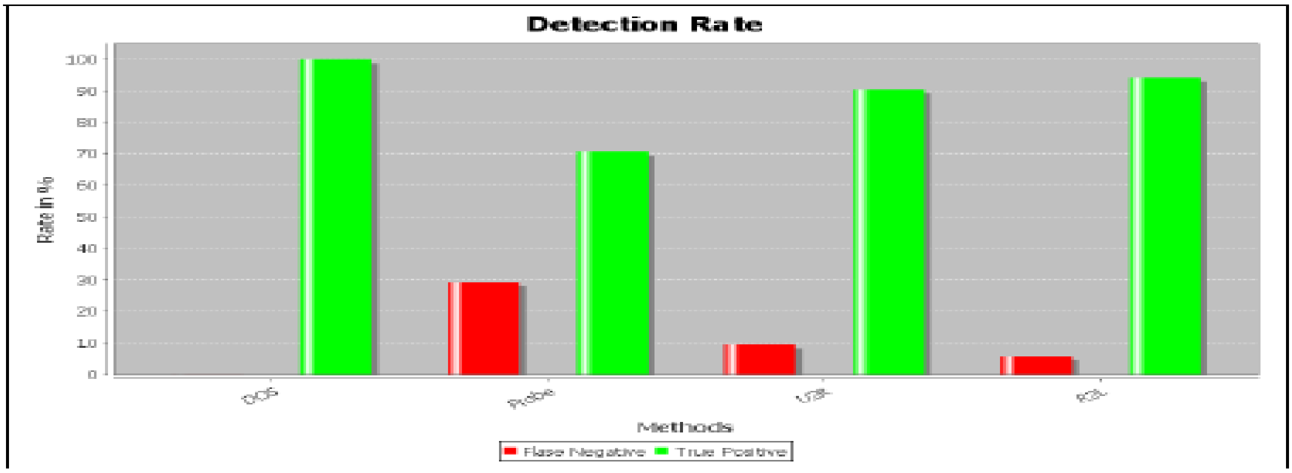


Figure 2: Naive Bayes, IBK, J48, Multilayer Perceptron, Bayes Net

Table 2: Results of Discretization Technique.

No.	Combinations	Accuracy	DOS	U2R	R2L	Probe
1	NB,BayesN,IBK,JRip,J48	99.38	99.90	63.63	79.45	68.50
2	BayesN,IBK,JRip,J48,PART	99.63	99.90	59.09	86.09	69.29
3	IBK,JRip,J48,PART,MLP	99.61	99.90	71.42	92.75	69.92
4	JRip,J48,PART,MLP,NB	99.67	99.76	60	84.23	69.21
5	J48,PART,MLP,NB,BayesN	99.38	99.88	59.09	79.45	68.94
6	PART,MLP,NB,BayesN,IBK	99.38	99.87	68.18	86.30	69.65
7	MLP,NB,BayesN,IBK,JRip	99.47	99.88	59.09	86.30	69.17
8	NB,IBK,J48,MLP,BayesN	99.38	99.89	63.63	80.82	68.98
9	BayesN,JRip,PART,NB,MLP	99.51	99.88	54.54	83.56	89.21
10	NB,BayesN,IBK,J48,PART	99.37	99.90	63.63	79.45	68.98
11	NB,BayesN,IBK,JRip,PART	99.51	99.87	54.54	83.56	69.02
12	NB,JRip,PART,MLP,IBK	99.55	99.89	68.18	83.56	69.89
13	BayesN,JRip,MLP,IBK,J48	99.60	99.88	59.09	87.67	69.37
14	NB,J48,IBK,MLP,JRip	99.55	99.89	68.18	86.30	69.93
15	NB,J48,IBK,MLP,PART	99.61	99.81	72.72	87.67	69.69
16	BayesN,JRip,PART,IBK,MLP	99.55	99.88	68.18	86.30	69.93
17	NB,JRip,MLP,BayesN,J48	99.45	99.90	61.90	84.05	68.73
18	NB,Ibk,PART,JRip,J48	99.60	99.90	66.66	88.42	69.36
19	Nb,J48,BayesN,JRip,PART	99.38	99.89	57.14	81.15	68.69
20	BayesN,IBK,PART,J48,MLP	99.60	99.89	71.42	92.75	69.64
21	BayesN,JRip,MLP,J48,PART	99.66	99.87	61.90	91.30	69.40

Discretization Results Analysis:

1. In discretization only one combination gave best results in all perspectives i.e.Naive Bayes,J48,IBK,Multilayer Perceptron, and PART

Accuracy=99.61
DoS=99.89
U2R=76.72
R2L=92.75
Probe=69.69

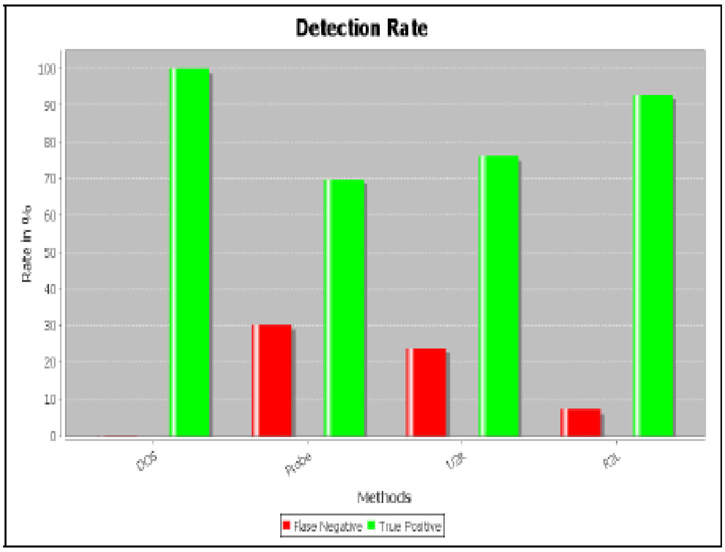


Figure 3: Multilayer perceptron, Naive Bayes, Bayes Net, IBk, JRip

VI.CONCLUSION AND FUTURE SCOPE

The purpose of these experiments was to study the performance analysis of pre-processing techniques using ensemble of 5 classifiers. Hence after performing the experiments it was observed that detection of U2R and R2L attacks was detected at a high rate by ensemble of classifiers. After the analysis of result the accuracy achieved by both pre-processing techniques are best. But Normalization pre-processing technique gave best results the Discretization for detection of U2R and R2L attack. The probe attack was also detected with better rate by Normalization. Dos attack was successfully detected by most of the combinations. Hence from the analysis of the results it can be concluded that Normalization pre-processing technique was more efficient then Discretization. Ensemble of 5 classifiers were more efficient than single classifiers. Similarly in future more experiments with respect to different techniques and classifiers must be performed.

REFERENCES

[1] V. D. Katkar and S. V. Kulkarni, "Experiments on detection of Denial of Service attacks using ensemble of classifiers," Green Computing, Communication and Conservation of Energ(ICGCE), 2013 International Conference on, Chennai, 2013, pp. 837-842.

[2] S. Choudhury and A. Bhowal, "Comparative analysis of machine Learning algorithms along with classifiers for network intrusion Detection," Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015 International Conference on, Chennai,2015,

[3] P. Sornsuwit and S. Jaiyen, "Intrusion detection model based Ensemble learning for U2R and R2L attacks," 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE), Chiang Mai, 2015, pp. 354-359.

[4] K. Elekar, M. M. Waghmare and A. Priyadarshi, "Use of rule base Data mining algorithm for intrusion detection," Pervasive Computing (ICPC), 2015 International Conference on, Pune, 2015, pp. 1-5.

[5] T. Garg and S. S. Khurana, "Comparison of classification Techniques for intrusion detection dataset using WEKA," Recent Advances and Innovations in Engineering (ICRAIE), 2014, Jaipur, 2014, pp. 1-5.

[6] H. Chauhan, V. Kumar, S. Pundir and E. S. Pilli, "A Comparative Study of Classification Techniques for Intrusion Detection ," Computational and Business Intelligence (ISCBI), 2013 International Symposium on, New Delhi, 2013, pp. 40-43.

[7] S. Roy, P. Krishna and S. Yenduri, "Analyzing Intrusion Detection System: An ensemble based stacking approach", 2014 IEEE s International Symposium on Signal Processing and Information Technology (ISSPIT), 2014.

[8] P. Amudha, S. Karthik and S. Sivakumari, "Intrusion detection Based on Core Vector Machine and ensemble classification Methods", 2015 International Conference on Soft-Computing and Networks Security (ICSNS), 2015.

[9] G. Nadiammai and M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian Informatics Journal, vol. 15, no. 1, pp. 37-50, 2014.

[10] F. Nia and M. Khalili, "An efficient modeling algorithm for intrusion detection system using C5 and Bayesian Network Structure", 2nd International conference of Knowledge based Engineering,2016.

[11] F. Nia and M. Khalili, "An efficient modeling algorithm for Intrusion detection systems using C5.0 and Bayesian Network Structures", 2015 2nd International Conference of Knowledge- Based Engineering and Innovations (KBEI), 2016