International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-6 Issue-3, February 2017

Ensemble of Classifiers for Intrusion Detection System

Sonali Kadam, Rutuja Pawar, Shweta Phule, Priyansha Kher, Manisha Kumari

Abstract- The continuous growth in Network attacks is being a serious problem in software industry. Intrusion detection framework is utilized to distinguish and break down system assaults so IDS should be upgraded that can screen the framework and can trigger the readiness in the framework. Numerous calculations have been proposed by various creators to enhance the execution of IDS yet at the same time they can't give appropriate or finish arrangement. In proposed framework creators perform probes distinctive blends of Bayesian system, Naïve Bayes, JRip, MLP, IBK, PART and J48 classifier. What's more for each mix two pre-processing procedures Normalization and discretization will be connected. The advantage of proposed approach is the combination detecting majority attacks will be ensemble with the respective pre-processing technique. Hence, any kind attack in the network can be detected with best accuracy.

Keywords: Bayesian network, Intrusion Detection System, IBK. JRip, J48, MLP, Naïve bayes, PART.

I. INTRODUCTION

Ensemble is one of the method where different classifiers are combined in order to achieve good results. In Intrusion Detection System the main use of Ensemble-classifier is to combine multiple classifiers to reach a more precise inference result than a single classifier and is noticed that accuracy of classifiers increases when combined together .As increase in the internet services and usage with open access to sensitive data, necessity of security to systems had become a need of the hour. Intrusion Detection Systems (IDSs) provide an important layer of security for computer systems and networks, and are becoming more and more crucial issue and is used to protect computer system from the risk of theft from intruders. In recent year internet security volume and sophisticated target network attacks has been increased substantially. There is increment in number of dangers and vulnerabilities like-business system framework, military and so on. This drives Intrusion Detection System as a noteworthy research range. Intrusion Detection System is isolated into two sections –Anomaly-based and Misuse-based. Peculiarity based model is utilized for the deviation of new information from the pre-characterized profile of information.

Manuscript published on 28 February 2017.
***** Correspondence Author (s)
Sonali Kadam, Bharati Vidhyapeeth's College of Engineering for Women, Pune (Maharashtra). India.
Rutuja Pawar, Bharati Vidhyapeeth's College of Engineering for Women, Pune (Maharashtra). India.
Shweta Phule, Bharati Vidhyapeeth's College of Engineering for Women, Pune (Maharashtra). India.
Priyansha Kher, Bharati Vidhyapeeth's College of Engineering for Women, Pune (Maharashtra). India.
Manisha Kumari, Bharati Vidhyapeeth's College of Engineering for Women, Pune (Maharashtra). India.

Abuse based is otherwise called learning based, is utilized to perform discovery through coordinating the new information with the known assaults in database. The essential requirement of any IDS is accuracy. The other requirements are extensibility and adaptability. The major problem with IDS is detection of false attacks. In this work, we intended to rank ten different supervised machine learning algorithms based on the metrics to evaluate their performance in classification. False positive rate, false negative rate, Precision, Recall and Detection accuracy are the metrics taken and there exists a trade-off among these metrics. Hence ensemble of classifier giving good accuracy is used for the proposed the system. The proposed idea will be used for Intrusion Detection System of a small organization for prevention of confidential data. It will help the organization by detecting the future attacks that harm the confidential data. Hence, ensemble of classifier subsequently investigates different group of Bayesian classifier, Naïve Bayes, IBk, JRip, MLP, J48 and PART will be performed.
Additionally paper is separated into five segments – Section II depicts the literature survey; Section III states the background work of the paper; Section IV portrays the classifiers in proposed framework; Section V proposed framework is depicted; Section VI the result of the project is examined. In area 7 conclusion and future degree is mentioned.

II. REVIEW OF LITERATURE SURVEY

In the research done by Vijay Katkar and Siddhart Kulkarni [1] after performing experiments the results stated that ensemble of J48, REPtree and Bayesian Network detected DoS attacks with significant accuracy.

In the research done by Sumoli Choudhary and Anir-ban Bhou-wan [2] No. of classifiers used to check for accuracy. Results proved that Bayes Net and Random Forest Tree are best classifica-tion techniques and boosting gives good result. In the research done by Kailas Elekar et. al.[4] The performance of all the classification techniques were evaluated on the basis of cross validation and test data from which, PART gave better results than others.
In the research done by Ploy phan Sornsuwit and Saichon Jaiyen[3] Authors accepted the use of correlation based algorithms for reduction of features in dataset..

In the research done by Tanya Garg and Surinder Singh Khurana [5] Garrett's Ranking Technique has been applied to rank different classifiers according to their performance. Rotation Forest classification approach performed the rest.
In the research done by Salah Eddine, Lalia Saoudi, Ourada Lounis [6] A genetic algorithm approach is presented to efficiently detect various types of attacks of network intrusions using NSL-KDD99 dataset.

In the research done by Himadri Chauhan, Vipin Kumar et al [6] selection of top 10 classifiers-SGD, IBK, JRip, PART, Random Forest, Random Tree, Logistics, Bayes Net were done. Out of those Random Forest, IBK gave better results with respect to accuracy and less processing time.
In the research done by Fengu Du, Zhejiang Coll. A new type of defense technology introduced as the way of attacks keep changing. It makes up some traditional security technologies such as information encryption, VPN, Firewall.

In the research done by T.Subbulakshmi, A.Ramamoorthi, Dr S.Mency Shailinis [9] Soft computing techniques applied to the intrusion detection system using the standard datasets (KDD99) to determine the Detection rate and False Positive rate. In the research done by Anazida Zainal, Mohd Aizaini Maroof, Ajith Abraham [10] Results show an improvement in detection accuracy for all classes of network attacks Probe,U2R, Normal, DoS etc.

The goal of this research work is to apply the two main preprocessing techniques normalization and discretization to the intrusion detection system using standard dataset. The standard dataset used is KDDCUP'99 dataset for Misuse detection system. And a constantly updated database is used to store the signature of different attacks like Normal, Probe, U2R, DoS etc. In this research five classifiers are ensemble to lower the error rate, to reduce over fitting of data and to taste great of attacks than single classifiers.

III. BACKGROUND

As of late, a group technique is boundlessly used to distinguish interruption 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 done. Initially many tests were performed on these grouping calculations and after that they were chosen. Out of those J48, IBK gave better outcomes concerning exactness and less handling time [6]. In 2015, Sumoli Choudhary and Anirban Bhouwan proposed a framework with a few characterizations systems and machine learning calculations. The characterization strategies utilized where Bayes Net, IBK, JRip, PART, Logistic, J48, Random Tree, J48 and REP tree. All these were outfit with machine calculations Boosting, Bagging and Stacking approach. Hence the outcomes demonstrated that Bayes Net and J48 Tree are best characterization strategies and boosting gives great outcome [2].In 2015 Kailas Elekar et.al analysed administer based order procedures –Decision Tree, PART, Zero R, One R, and JRip. The execution of all these characterization strategies were assessed on the premise of cross approval and test information from which, PART gave preferred outcomes over others [4]. In 2013Vijay Katkar and Siddhart Kulkarni proposed a framework for recognition of Denial of Service assault in system. The classifiers utilized for analyses were Naïve Bayesian, Bayesian Network, Sequential Minimal Optimization, J48, choice tree. Subsequent to performing tests the outcomes expressed that outfit of J48, REP tree and Bayesian Network distinguished DoS assaults with huge exactness. The creators additionally demonstrated that group of classifier can be utilized as opposed to building new classifier [1]. In 2015 creators Ployphan Sornsuwit and Saichon Jaiyen proposed an Intrusion location based model for identification of U2R and R2L assaults. The creators embraced Adaboost calculation for troupe of powerless learners and to support the execution. The frail learners utilized were Decision Tree, SVM, Naïve Bayes and MLP.In expansion to this, creators acknowledged the utilization of relationship based calculations for decrease of elements in dataset.The comes about demonstrated that gathering of Naïve Bayes and MLP gave great outcomes for U2R and R2L assault with most noteworthy sensitivity. Decision Tree flopped in thiscase [3].In 2014 Tanya Garg and Surinder Singh Khurana performed correlation of various arrangement procedures for Intrusion Detection System. Creators utilized Garette positioning technique to rank the classifiers. As per Garette positioning Rotation Forest was positioned 1[5].

IV. THEORETICAL PRELIMINARIES

A. Training Dataset:

To train the classifier we are utilizing 'training data set index'. A preparation set is an arrangement of information used to find conceivably prescient connections. After the preparation is finished the informational collection is handled for the approval part so to assess the model properties.

B. Pre-processing Techniques:

1. Normalization:

Normalization is one of the pre-processing technique which scales up and scales down the data i.e. manipulation of data is done before it is used in further stages. Normalization has many techniques like Min-Max normalization, Z-score normalization, Integer Scaling Normalization and Decimal scaling normalization. In min-max relationship is kept up between unique information and it gives straight change. Primary point of utilizing this method is that it deciphers the outcome precisely, expels deviations and anomalies, uncover the examples, and so forth.

2. Discretization:

Discretization is used as a pre-planning wander for machine learning counts that handle simply discrete data. Discretization also goes about as a component decision technique that can out and out influence the execution of request figuring's which is used as a piece of the examination of high-dimensional biomedical data. It has basic repercussions for the examination of high dimensional genomic and proteomic data. It is the route toward changing a constant regarded variable into a discrete one by making a course of action of touching intervals or similarly a game plan of cut concentrations that clear up the extent of the variable's qualities. Discretization methodologies fall into two unmistakable characterizations: unsupervised, which don't use any information in the target variable and coordinated systems, which do. It has been measured that controlled discretization is more worthwhile to portrayal than unsupervised discretization. Coordinated discretization methodologies will discretize a variable to a single interval if the variable has for all intents and purposes no association with the goal variable. This enough clears the variable as a commitment to the gathering estimation.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

C. Classifiers:

1. Naïve Bayes:

Naïve Bayes is a simple technique which assigns class labels to problem instances. Naïve Bayes is one of the classifier in which every feature has its own independent value which is not dependent on any other feature. Naive Bayes classifier can get rid of large scale classification problems even if the entire training set is not fitted in the memory. Advantages of naïve Bayes are that it is best for spam filtering and document classification.

2. J48:

J48 is an Open Source use of C4.5 strategy gave in the Weka interface. C4.5 technique creates a decision tree. C4.5 is a growth of the ID3 classifier. The tree which is created by C4.5 figuring is used for course of action, which is known as a true classifier. C4.5 estimation uses tantamount procedure to make decision tree, except for the usage of information entropy. At each centre point of the tree, C4.5 picks the nature of the data that most suitably parts its course of action of tests into subsets enhanced in one class or the other. With the ultimate objective of settling on choice property which has the most amazing institutionalized information get is picked.

3. IBK:

Essentially "IB" remains for Instance-Based and "k" determines number of neighbours that are inspected. IBK actualizes k-Nearest Neighbour calculation. In IBK, information is spoken to in a vector space. It is utilized for grouping, relapse and evaluating constant factors. In light of cross approval it can choose suitable estimation of K. Separate weighting should likewise be possible.

4. JRip:

It is one of the fundamental and most famous calculations. Every one of the classes are analysed as the developing size and the underlying arrangement of standards are produced by utilizing the decreased mistake rate. In JRip every one of the cases continues by a specific choice in training information as a class, and it finds that cover every one of the individuals from a class. After that it continues, to the following class and this method goes ahead till the end, until all classes are secured legitimately

5. PART:

It is known as the different and-Conquer administer learner. It creates all the arrangement of tenets called as 'Choice rundown'. As the principles are created, another arrangement of information is contrasted with each control and after that the things are doled out to the class of first coordinating standard. It constructs a halfway c4.5 choice tree in each cycle and makes the best leaf into a run the show.

6. MLP:

Multi-Layer Perceptron (MLP) is an arrangement of clear neurons which is called as perceptron. MLP is a feed forward neural framework with no less than one layers among data and yield layer. Feed forward suggests that data streams in one course from commitment to yield layer (forward heading). This sort of framework is set up with the back inducing learning computation. MLPs are extensively used for instance affirmation, game plan, figure and gauge. The issues which are not directly distinguishable can be settled by Multi-Layer Perceptron. The perceptron processes a solitary yield from different genuine esteemed sources of info. It figures yield by framing a straight blend as indicated by its info weights. The quantity of concealed units can likewise be indicated.

7. Bayes Network:

Bayes net (Bayesian Network) is a classifier based on probabilistic model. It represents a set of random variables with their conditional dependencies through a directed acyclic graph (DAG). It uses various search algorithms and quality measures. Bayes net only relates that particular node which is probabilistically related by some casual dependencies which gives a huge saving of computation. There is no need to store all the possible configuration of state, the only thing has needed to relate with all possible related combination of sets. This makes a huge saving of computation and space table. Another reason to use Bayes net is that they are adaptable. It starts with a small and limited knowledge about a domain and acquire new knowledge. So, one's need not have to keep complete knowledge about the instance or domain. Advantage of using Bayes net is that probabilities need not to be exact. It should be approximate probabilities. It uses casual conditional probabilities than reverse to estimate. Because it is better to estimate probabilities "in the forward direction".

D. Test Dataset:

Test set is the informational index on which we apply our model and check whether it is working accurately and yielding expected and coveted outcomes or not. Test set resembles a test to the model. Testing information is the information, whose result is as of now known and is utilized to decide the exactness of the machine learning calculation, in light of the preparation information.

V. PROPOSED SYSTEM

Adoption of 7 different classifiers- Bayesian Network, Navies bayes, PART, JRip, MLP, J48 and IBK is being done in the experiments. Experiments on different combinations of classifiers will be tested for analysing attacks. According to survey accuracy increases after data cleaning or data preprocessing. Hence, for better accuracy and results two preprocessing techniques – Normalization and Discretization will be performed on training data set and test data .Combinations of 5 classifiers will be performed. For each combination accuracy, precision and false alarm rate will be calculated. Depending on the survey, selection of classifiers for various attacks is being done. As ensemble of classifiers increases the accuracy, the classifiers having good accuracy are used. The main advantage of proposed idea is, implementation of such classifiers will be done which can detect all the attacks. According to different author's ideas and results different classifier are responsible for detection of different attacks. Expansion to these pre-preparing procedures will contribute for more exact outcomes. Subsequently in proposed thought group of such classifiers will be tried with pre-handling methods for the best outcomes. Exactness is one of the one of the vital variable which will be tried while investigates distinctive mixes of classifiers.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The proposed framework will be executed with the point of recognizing distinctive assaults like Normal, Probe, U2R, R2L, DoS. Consequently recognizing all assaults in system will help in counteractive action of various unlawful practices. Subsequently it will be helpful in Intrusion Detection Systems.

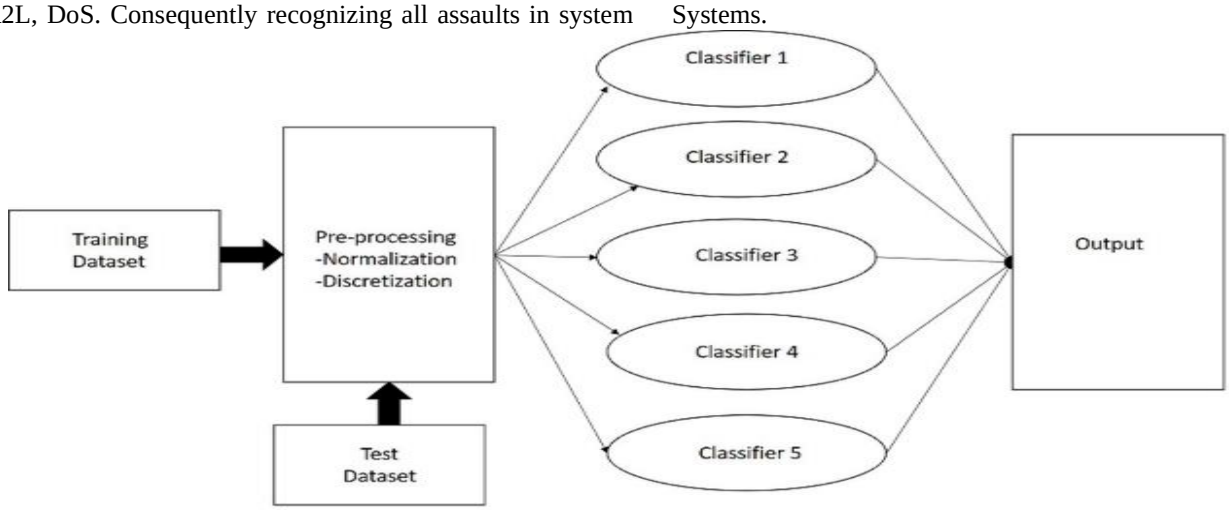


Fig. 1. Ensemble of classifier

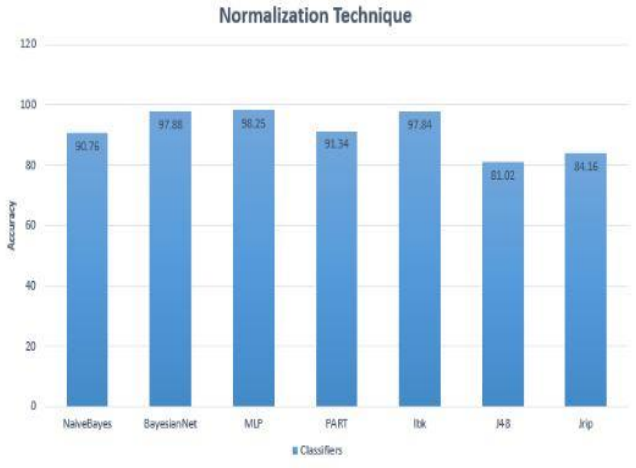
VI. RESULT

Single accuracy using Normalization:

Classifiers	DoS	U2R	R2L	Probe
Naïve Bayes	1937774	14	254	14478
Bayesian Net	1937686	11	571	19800
MLP	1938995	12	560	19928
PART	1936880	6	363	19743
IBk	1938172	14	557	20021
J48	1938733	12	69	19737
JRip	1937144	0	555	14936

Normalization table

Using normalization technique, authors detected Dos, U2R, R2L, Probe attacks. Author observed that Multilayer Perceptron detects more DoS attacks compared to others. While PART detects at least DoS attacks. Naive Bayes and IBk detects same amount of U2R attacks. Bayesian Network detects R2L attacks for efficiently. And IBk has more R2L attacks detected compared to others



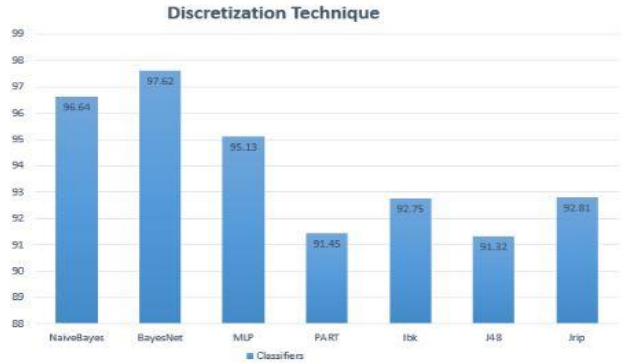
In this, author has plotted a graph with classifiers on X-axis and Accuracy on Y-axis showing the relation between these two attributes for normalization technique. In this we can observe that Multilayer Perceptron gives highest accuracy compared with others. IBk and Bayesian Network almost have the same accuracy.

Single accuracy using Discretization:

Classifiers	DoS	U2R	R2L	Probe
Naïve Bayes	1938172	14	557	19021
Bayesian Net	1937686	11	571	19800
MLP	1940655	23	517	19389
PART	1919693	07	331	18714
IBk	1940887	22	519	19552
J48	1909744	07	517	15121
JRip	1938025	04	369	18854

Discretization

In this, author observed that Multilayer Perceptron detected more DoS attacks. Bayesian Network has detected highest R2L attacks and Probe attacks.



In this, author has plotted a graph for discretization technique. Bayesian Network has the highest accuracy than others and is observed as best for discretization technique. Multilayer Perceptron has less accuracy compared to normalization technique.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Naive Bayes has moderate accuracy compared to other. Author has ensemble five classifiers out of seven using normalization technique. In which author calculated accuracies of 21 combinations from which ensemble of J48, PART, MLP, IBk and Bayesian Network gave highest accuracy. Ensemble of classifiers also detected DoS attack, U2R attack, R2L attack and Probe attack. The table contains ensemble of classifiers which give best results for at least one of the attacks and one which gives the highest accuracy of all. Similarly, author is going to calculate the accuracies of ensemble of classifiers using discretization technique and observe which combination detects the attack accurately and efficiently.

Ensemble of Classifiers	Accuracy	DoS	U2R	R2L	Probe
J48,PART,IBk,MLP, BayesNet	98.39	1939528	14	560	20035
BayesNet,NB, J48,PART,MLP	98.27	1940710	14	561	19685
IBk,NB,BayesNet, MLP,JRip	98.31	1939382	16	562	19935
JRip,BayesNet, MLP,PART,IBk	92.88	1940672	12	566	19804

Ensemble of classifiers using normalization technique

NB=Naïve Bayes
BayesNet=Bayesian Network

VII. CONCLUSION AND FUTURE SCOPE

The purpose of these experiments was to study the performance analysis of pre-processing techniques using ensemble of 5 classifiers. The accuracy of individual classifiers achieved better accuracies but detection rate of attacks were not achieved. Hence after performing the experiments it was observed that detection of U2R and R2L attacks was detected at a high rate by ensemble of JRip, Bayes Net, MLP, PART, IBk and IBk, Naive Bayes, Bayes Net, MLP, JRip. The rate of detection of attacks achieved were more than the individual accuracies of classifiers. Hence ensemble of 5 classifiers were more efficient than single classifiers. Similarly more experiments with respect to different techniques and classifiers must be performed.

REFERENCES

1. V. D. Katkar , S. V. Kulkarni, "Experiments on detection of Denial of Service attacks using ensemble of classifiers, *Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on*, Chennai, 2013, pp. 837-842.
2. S. Choudhury, 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, pp. 89-95.
3. P. Sornsuwit , 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 , 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 and 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. 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.
8. G. Nadiammai , M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian Informatics Journal, vol. 15, no. 1, pp. 37-50, 2014.
9. F. Nia , M. Khalili, "An efficient modelling algorithm for intrusion detection systems using C5.0 and Bayesian Network structures", 2015 2nd International Conference of Knowledge-Based Engineering and Innovations (KBEI).

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)