JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Experimental Analysis of k-Nearest Neighbor, Decision Tree, Naive Baye, Support Vector Machine, Logistic Regression and Random Forest Classifiers with Combined Classifier Approach for NIDS

Nilesh B. Nanda[1*] , Ajay Parikh[2]

[1]Research Scholar (Computer Science) Gujarat Vidyapith, Ahmadabad-Gujarat, India
[2]Head Department of Computer Science Gujarat Vidyapith Ahmedabad-Gujarat, India

[*]*Corresponding Author: nilideas@yahoo.co.in*

Available online at: www.ijcseonline.org

Accepted: 25/Sept./2018, Published: 30/Sept./2018

Abstract—In traditional studies about the classification, there are three non-parametric classifiers, Random Forest (RF), k-Nearest Neighbor (kNN), and Support Vector Machine (SVM), has been said as the most classifiers at producing excessive accuracies. In this study, Tested and Compared the performances of the kNN, Naïve Baye, Decision Tree, Support Vector Machine, Random Forest, Logistic Regression and Combined model over DOS and Normal attacks. These algorithms are among the most influential data mining algorithms in the research community. The detection of fraudulent attacks is considered as a classification problem. In this experiments have performed on different classification methods with the hybrid model on KDDCup99 Dataset. Here compared classifiers using models accuracy with confusion matrix. Cross-Validation means score used for efficiency. For this experiments used python and R programming for implementation. The different types of attacks are routine, DoS, Probe attacks, R2L, and U2R attacks.

Keywords— Network intrusion, support vector machine, decision tree, Decision Tree, detection.

I. Introduction

Data organization is very critical Classification is one of the most tedious jobs in data mining. In data classification, a classifier was determined from a set of training cases with class labels, and an instance often expressed by a tuple of attributes, where denotes the value of the attribute. Its classification accuracy or confusion matrix typically cover the act and effect of a classifier. Mostly network intrusions are the disturb of information security rules. At first, NIDS implemented for computer-based that located in the data center to examine the internal interfaces [1][2][3], but with the evolution of computer networks, the focus gradually shifted toward network-based. Network intrusion detection system (NIDS) performs packet logging, real-time traffic analysis of IP network, and tries to discover if an intruder is attempting to break into the system [4][5][6][10][11][12].

Different Attacks on the network can be referred to as Intrusion. Intrusion means any set of fake activities that attempt to leak the security standards of the information. Network Intrusion detection is one of the enormous information security problems. NIDS (Network Intrusion Detection System) assist the host in resisting internal and external network attacks[1]. In our research work, naive Bayesian classifier, support vector machine, decision tree, random forest, K-Nearest Neighbor, logistic regression, decision tree and voting classifier (combined algorithm) are presented based on a comprehensive analysis for the current research challenges in network intrusion detection. A new learning algorithm for adaptive network intrusion detection, which can handle the above-mentioned challenging issues. In this paper, we address some difficulties including data mining such as managing continuous attribute, dealing with missing attribute values, and decreasing noise in training data. This classifier will be evaluated on the NSL KDD dataset to identify attacks on the various attacks categories: Probe (information gathering), DoS (denial of service), U2R (user to root) and R2L (remote to local). The classifier's results are computed for comparison of feature reduction methods to show that the hybrid model is more efficient for network intrusion detection.[10][11][13].

In this work, we organized as follows. Section I gives Introduction. Section II discusses the literature survey. Section III overviews the intrusion detection system and classification. Section IV offers various data mining techniques and tools for NIDS. Section V presents the multiple datasets that are used to build a NIDS, and the next section is in conclusion.

II. Background

Various algorithms have been used in the security area and machine-based learning methods. In this paper, compare the very famous mining algorithm like SVM, Navie Baye, decision tree, KNN and combined model using with KDDCup train dataset[10][11].

The SVM is the best learning type of pattern algorithm for binary classification. It has applied to information security for network intrusion detection.

Decision tree and Naive Bayes techniques are used to automatically learn intrusion signatures, pattern and perform the classification activities in computer network systems as usual or intrusive[13-14].

K-mean clustering was used to perform importance features extraction through grouping over data and in unsupervised manner cluster the whole KDD cup'90 dataset into parts.

The voting classifier is a classification objective to link similar or conceptually designed machine learning classifiers to most multiple classifications and implements "hard" and "soft" grades. In difficult categories, predict the final class label as the class label that classification models have predicted more frequently. In the soft grade, predict class labels by calculating the average probability of the class. The main advantage is to provide excellent accuracy, speed and real-time sensing of intrusions. It also can update training and signature pattern dynamically.

III. NIDS DIFFERENT TYPE OF ATTACKS

3.1 **Probe attacks:** In the probing attack hacker scans system or network device and determine the weaknesses or vulnerabilities of infrastructure for exploited the system. This technique commonly used in data mining, e.g., saint, port sweep, mscan, nmap, etc.[7]

3.2 **DoS Attacks**: A DoS attack the hacker executes to make resource computer system network are too busy.

3.3. **Remote to User Attacks (R2L):** Attacker tries to gain access to remote machine because they do not have rights to access or does not have control of same.

3.4 **U2R Attacks**: Probe: Attacker tries to get information from the remote host without knowing actual users.

IV. KDD CUP'90 Dataset

The primary purpose of the KDD dataset has to use a simulation of a military network to consisting of three target various running machines, operating systems, and traffic. Duration of this simulate is a few weeks. Normal TCP connections are used to create a profile that expected in a military network for attacks identification.

Accuracy of Classification
Classification Accuracy.

$$\text{ACCURACY} = \frac{TP + TN}{TP + TN + FN + FP}$$

Recall :

$$\text{RECALL} = \frac{TP}{TP + FN}$$

Precision :

$$\text{ACCURACY} = \frac{TP}{TP + FP}$$

F-measure:

$$\text{F measure} = \frac{2 * Recall * Precision}{Recall + Precision}$$

V. EXPERIMENTAL RESULTS

In the experiments, used standard NSL-KDD dataset[10][13][14]:
1) redundant record is removed from the train set to eliminate the bias to the most frequent records using R Programming. The kddcup99 dataset used in this research of which 80% is treated as training data, and 20% is considered as testing data.
2) In this dataset, we have used 43 attributes for each connection record including class label containing attack types. Train set dimension: 395217 rows and Test set size: 98804 rows.
3) Duplicate records in test sets are removed using R programming.
4) The number of records in the test and train datasets is reasonable.
5) For feature selection, we used the random forest classifier and utilized ten attributes[9].

For the experiment, subsets of training and test dataset were utilized. In [21] the NSL-KDD'99 dataset is analyzed using all experimental algorithm. The dataset was clustered into normal, DoS, Probe, R2L, and U2R attacks. The proposed method is implemented by the R Programming, Python, Jupyter notebook with Anaconda Navigator software and tested on NSL-KDD dataset. The number of training and testing datasets which are used for the experiments are shown in Tables 4.1 and Graph.

Table 4.1. For the Experiments the number of training and test datasets used.

	Attack Class	**Frequency Percent Train**	**Attack Class**	**Frequency Percent Test**
DoS	312251	79.01	79207	80.17

Probe	3796	0.96	311	0.31
R2L	1125	0.28	1	0.00
U2R	35	0.01	17	0.02
Normal	78010	19.74	19268	19.50

Scenario 1: Training datasets used for the algorithm. Thus the training and test datasets are entirely separated from each other. Scenario 2: In training not only train dataset used but also a subset of the test dataset used. Thus the test and training datasets are not entirely separated from each other.

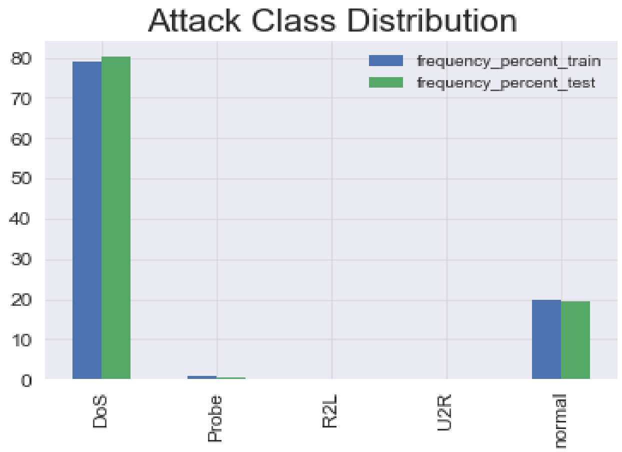

Fig: 1 Attack class bar plot

4.2 The simulated analysis of the NIDS methods of all classifier done by using well define performance measuring parameters[10][11]. Here, table 4.2 shows accuracy result of the Evaluates models and Test models using SVM, Decision tree and KNN algorithms. After analysis, it was found that the overall accuracy rate for Evaluates method of SVM is about 99.82% whereas the Test models are 99.94%. Decision tree accuracy is 100% during evaluates models and 99.83% during Test models. KNN algorithm accuracy becomes 99.99% whereas in test models 99.98%. So it is concluded that Evaluates models generate a more accurate result for network intrusion detection as compared to the Test method.

Table 4.2 Comparison of an accuracy rate of Evaluates models and test models with DOS attacks of classifiers model.

Accuracy		
Models	**Evaluates Models**	**Test Models**
SVM	0.998680548661	0.742272189944
Naïve Baye	0.864207000138	0.719434593425
Decision Tree	1.0	0.079076435524
KNN Model	0.999979183413	0.989184838653
LogisticRegression	0.99748279429	0.804270731155
VotingClassifier	0.999985588517	0.74575567796

Table 4.3 Comparison of Cross-Validation Mean Score of Evaluates models and test models with DOS attacks of SVM, Decision Tree, Naive baye, KNN model, Logistic regression and voting classifier.

Cross-Validation Mean Score	
Models	**Evaluates Models**
SVM	0.995542123111
Naïve Baye	0.862687892606
Decision Tree	0.996541343034
KNN Model	0.99870456924
LogisticRegression	0.996751029518
VotingClassifier	0.998321883536

Here, table 4.3 also shows the cross-validation mean score of evaluates the model for SVM, Naïve Baye, Random forest, KNN and Decision tree algorithm which shows that KNN gives good results compare to Decision tree and SVM.

VI. CONCLUSION

In this paper compare and analysis of a various model like SVM, Decision tree, Naïve Baye, KNN, decision tree, Logistic Regression and combined models for improving the NIDS and observing conclude that the performance of the hybrid model has significantly improved the classification accuracy, and it proves the importance of preprocessing in NIDS. As compared to the existing methods, Evaluates model fairly enhances the random forest classification accuracy of Dos attacks. Random forest classifier used to attributes selection to improve to the accuracy of the results. Hence conclude that the combined model of classifier proves to be an efficient classifier for DoS attacks. Using combined models like knn and support vector machine or decision Tree and Naive Baye and other technique which is a future work to be proposed to improve the efficiency of network intrusion detection system using Random Forest Classifiers for feature selection with combined classifiers.

REFERENCES

[1] D.Dennin, “An intrusion-detection model”, IEEE Transactions on Software Engineering, 2007.

[2] J. Frank,“Machine learning and intrusion detection: Current and future directions,” in Proceedings of the National 17th Computer Security Conference, Washington, D.C., 2014.

[3] Lee, W., Stolfo, S., &Mok, K. “A Data Mining Framework for Building Intrusion Detection Model.Proc”. IEEE Symp. Security and Privacy, 120-132, 1999.

[4] Amor, N. B., Benferhat, S., &Elouedi, Z., “Naive Bayes vs. Decision Trees in Intrusion Detection Systems.Proc.” ACM Symp.Applied Computing, 420- 424, 2014.

[5] Mukkamala, S., Janoski, G., &Sung, A., “Intrusion detection using neural networks and support vector machines,” Paper presented at

the International Joint Conference. On Neural Networks (IJCNN), 2012.

[6] Heba F. Eid, Ashraf Darwish, Aboul Ella Hassanien, and Ajith Abraham, "Principal Components Analysis and Support Vector Machine based Intrusion Detection System," IEEE, 2017.

[7] T.Shon, Y. Kim, C.Lee and J.Moon, "A Machine Learning Framework for Network Anomaly Detection using SVM and GA", Proceedings of the 2015 IEEE, 2015.

[8] KyawThetKhaing (2010), Recursive Feature Elimination (RFE) and k-Nearest Neighbor (KNN) in SVM.

[9] H. Liu and H. Motoda(1998), Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic.

[10] N. Nanda, A. Parikh,"Classification and Technical Analysis of Network Intrusion Detection Systems," International Journal of Advanced Research in Computer Science, Volume 8, No. 4, May-June 2017.

[11] N. Nanda, A. Parikh, "Network Intrusion Detection System: Classification, Techniques and Datasets to Implement," International Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454-4248 Volume: 4 Issue: 3 106 – 109,2018.

[12] P. Tembhare, N. Shukla. "An Integrated and Improved Scheme for Efficient Intrusion Detection in Cloud", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.3, pp.74-78, June 2017.

[13] P. Dehariya, "An Artificial Immune System and Neural Network to Improve the Detection Rate in Intrusion Detection System", International Journal of Scientific Research in Network Security and Communication, Volume-4, Issue-1, Feb- 2016.

Author Profile

Nilesh Nanda pursed M.Phil of Computer Science from Gujarat Vidyapith, Ahmedabad, India in 2013 and currently pursuing Ph. D of Computer Science in the field of Network intrusion detection system from Gujarat Vidyapith. He is currently working as Computer Programmer in VVP Engineering College, Rajkot Gujarat (INDIA) Since 2000.

Dr. Ajay Parikh, Professor & Head, Department of Computer Science, Gujarat Vidyapith, Ahmedabad (Gujarat) INDIA. He pursued Master of Science from Gujarat University. He pursued M. Phil and Ph.D. of Computer from Gujarat Vidyapith. He has published 11 research papers in various conferences (National & international) has an excellent academic line of experience and published. His area of interest Machine Learning, Data Science, SOA, ICT Application in animal health care and Rural Development. He guided many Ph. D. and M.Phil Scholar students. He organized and participated in various seminars, Workshop, and training camps. He delivered lectures in refresher courses. He delivered Radio and TV talk. He delivered international and national lectures. He is a member of organizing a committee, program committee, international conference, workshop and reviewer in journals. He reviewed the research project. He has membership in professional and other bodies.