A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection

S. Revathi
Ph.D. Research Scholar
PG and Research, Department of Computer Science
Government Arts College
Coimbatore-18, India

Dr. A. Malathi
Assistant Professor
PG and Research, Department of Computer Science
Government Arts College
Coimbatore-18, India

Abstract

During the last decade the analysis of intrusion detection has become very important, the researcher focuses on various dataset to improve system accuracy and to reduce false positive rate based on DAPRA 98 and later the updated version as KDD cup 99 dataset which shows some statistical issues, it degrades the evaluation of anomaly detection that affects the performance of the security analysis which leads to the replacement of KDD dataset to NSL-KDD dataset. This paper focus on detailed study on NSL- KDD dataset that contains only selected record. This selected dataset provide a good analysis on various machine learning techniques for intrusion detection.

Keyword: NSL-KDD, Data Mining Technique and KDD Cup 99

I. Introduction

With the colossal growth of computer network all the computer suffers from security vulnerabilities which are difficult and costly to be solved by manufactures [1]. There is no disputing fact that the number of hacking and intrusion incidents is increasing year to year as technology rolls out, unfortunately in todays interconnected Ecommerce world there is no hiding place. The research in intrusion detection mainly based on misuse or anomaly detection in which misuse generally favored in commercial use. The anomaly detection fully based on theoretical methods for addressing novel attacks.

The research analysis for anomaly detection fully based on several machine learning methods on various training and testing dataset [2]. Our study analysis the inherent problem in KDDcup 99 dataset and the solution as study of NSL-KDD dataset for finding accuracy in intrusion detection. The first important deficiency in the KDD [3] data set is the huge number of redundant record for about 78% and 75% are duplicated in the train and test set, respectively. Which makes the learning algorithm biased, that makes U2R more harmful to network. To solve these issues a new version of KDD dataset, NSL-KDD is publicly available for researchers through our website. Although, the data set still suffers from some of the problems discussed by McHugh [4] and may not be a perfect representative of existing real networks, because of the lack of public data sets for network-based IDSs, we believe it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods.

In this paper, we have provided a solution to solve the issues, resulting in new train and test sets which consist of only selected records from complete KDD data set that does not suffer from any of the mentioned problems. Further, the number of records in the train and test sets are limited. This advantage makes it reasonable to run the experiments on using complete dataset without small portion. Therefore, the evaluation results of different research work will be consistent and comparable.

The rest of the paper is structured as follows: section II present some related work based on intrusion detection research. Section III explains detailed description of the attacks present in NSL-KDD dataset. Section IV summarize in detail about analysis of NSL KDD dataset on various data mining technique. Section V explain the experimental analyses on various attacks using different machine learning techniques. The conclusion and future work is summarized in section VI.

II. Related Work

The inherent problem of KDD dataset leads to new version of NSL KDD dataset that are mentioned in [6, 7]. It is very difficult to signify existing original networks, but still it can be applied as an effective benchmark data set for researchers to compare different intrusion detection methods [4]. In [7] they have conducted a statistical analysis on this data set and found two important issues which highly affect the performance of evaluated system, and results in very poor evaluation of anomaly detection approaches. To solve these issues, they proposed a new dataset, NSL-KDD, which consists of only selected records form the complete KDD dataset and does not suffer from any of the mentioned shortcomings.

In [5] they use k mean clustering technique on NSLKDD dataset to find the accuracy for intrusion detection. Shilpa et.al [8] used principal component analysis on NSL KDD dataset for feature selection and dimension reduction technique for analysis on anomaly detection. Generally, Data mining and machine learning technology has been widely applied in network intrusion detection and prevention system by discovering user behavior patterns from the network traffic data.

III. Dataset Description

The statistical analysis showed that there are important issues in the data set which highly affects the performance of the systems, and results in a very poor estimation of anomaly detection approaches. To solve these issues, a new data set as, NSL-KDD [6] is proposed, which consists of selected records of the complete KDD data set. The advantage of NSL KDD dataset are

1. No redundant records in the train set, so the classifier will not produce any biased result
2. No duplicate record in the test set which have better reduction rates.
3. The number of selected records from each difficult level group is inversely proportional to the percentage of records in the original KDD data set.

The training dataset is made up of 21 different attacks out of the 37 present in the test dataset. The known attack types are those present in the training dataset while the novel attacks are the additional attacks in the test dataset i.e. not available in the training datasets. The attack types are grouped into four categories: DoS, Probe, U2R and R2L. Table 1 shows the major attacks in both training and testing dataset [5].

Table I Attacks in Testing Dataset

Attacks in Dataset	Attack Type (37)
DOS	Back,Land,Neptune,Pod,Smurf, Teardrop,Mailbomb,Processtable,Udpstorm,Apache2,Worm
Probe	Satan,IPsweep,Nmap,Portsweep,Mscan,Saint
R2L	Guess_password,Ftp_write,Imap,Phf,Multihop,Warezmaster,Xlock,Xsnoop,Snmpguess,Snmpgetattack,Httptunnel,Sendmail, Named
U2R	Buffer_overflow,Loadmodule,Rootkit,Perl ,Sqlattack,Xterm,Ps

Fig 1 and 2 explains about the analysis of NSL KDD dataset in detail and shows the number of individual records in four types of attacks for both training and testing.

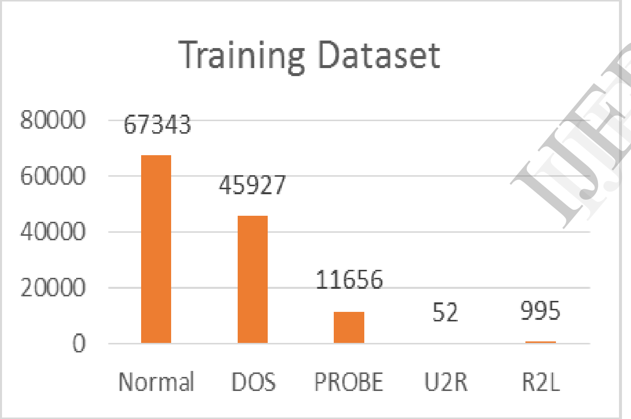

Fig 1. Number of Instance in Training Dataset

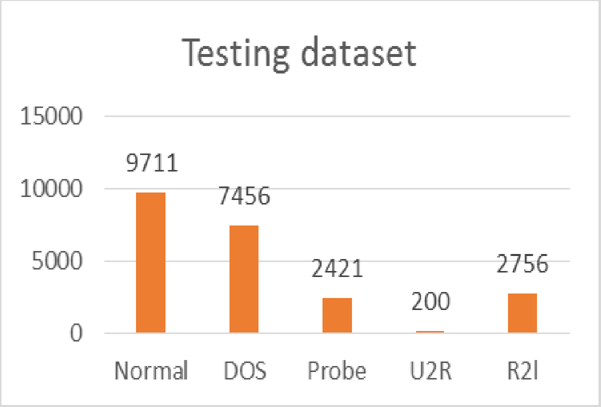

Fig 2. Number of Instance in Testing Dataset

IV. Data Mining Techniques:

Using data processing techniques, it perceive and extrapolate knowledge that may scale back the probabilities of fraud detection [9], improve audit reactions to potential business changes, and make sure that risks area unit managed in exceedingly a lot of timely and active manner. Additionally to employing a specific data processing tool, internal auditors will choose between a ranges of knowledge mining techniques. The foremost unremarkably used techniques embody artificial neural networks, decision trees, and nearest-neighbor methodology. Each of the techniques are analyzed the knowledge in numerous ways:

- Artificial neural networks are unit non-linear, predictive models that learn through training. Though they're powerful predictive modeling techniques. The auditors will simply use them is reviewing records to spot fraud and fraud-like actions, they're higher utilized in things wherever they will be used and reused, like reviewing MasterCard transactions each month to envision for anomalies.

- Decision trees are unit arborous structures that represent decision sets. These choices generate rules that are used to classify data.

- The nearest-neighbor methodology classifies knowledge set records supported similar data in an exceedingly historical dataset. Auditors will use this approach to outline a document that's fascinating to them and raise the system to go looking for similar things.

Each of these approaches has both advantages and disadvantages that need to be considered prior to their use. Neural networks, which are difficult to implement, require all input and resultant output to be expressed numerically, thus needing some sort of interpretation. The decision tree technique is the most commonly used methodology, because it is simple and straightforward to implement and the nearest-neighbor method relies more on linking similar items. A good way to apply advanced data mining techniques is to have a flexible and interactive data mining tool that extract, import, and analyze the data. On integrating data mining with warehouse it simplifies mining result.

Irrespective of good anomaly detection methods are used, the problems such as high false alarm rates is difficult in finding proper features, and high performance requirements still exist. Therefore, if we are able to mix the advantages of learning schemes in machine learning methods, according to their characteristics in the problem domain, then the combined approach can be used as an efficient means for detecting anomalous attacks. Some of the classification algorithm that most commonly used to classify the dataset are SVM, J48, Random forest, CART and Navie Bayes [10].

V. Experimental Result and Analysis

The data in NSL-KDD dataset is either labeled as normal or as one of the 24 different kinds of attack. These 24 attacks can be grouped into four classes: Probe, DoS, R2L, and U2R. The effectiveness of the algorithm is performed in weka tool [11]. It is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [12]. WEKA consists of four application namely Explorer, Experimenter, Knowledge flow, Simple Command Line Interface and also Java interface. The experimental steps are as follows

1. Select and preprocess the dataset.
2. Run the classifier algorithm.
3. Compare the classifier result.

The first step is to perform discretization as preprocess. Discretization is the process of turning numeric attributes into nominal attributes. The main benefit is that some classifiers can only take nominal attributes as input, not numeric attributes. Another advantage is that some classifiers that can take numeric attributes can achieve improved accuracy if the data is discretized prior to learning. From 41 attribute we have filtered to 13 feature vectors by using CFS subset technique to get an optimum selection from complete dataset for training as well as for testing experiments. Table II shows the test accuracy that achieved by using the six algorithms for the full dimension data and also after the feature reduction with CFS subset technique his shows that CFS subset can be used with any classification algorithms without much reduction in the test accuracy.

Table II. Test Accuracy for different classes of attacks

Classification Algorithm	Class Name	Test Accuracy (%) with 41 Features	Test Accuracy (%) with 15 Features
Random Forest	Normal	99.1	99.8
	DOS	98.7	99.1
	Probe	97.6	98.9
	U2R	97.5	98.7
	R2L	96.8	97.9
J48	Normal	78.9	87.5
	DOS	82.4	88.3
	Probe	80.2	86.0
	U2R	73.9	75.5
	R2L	87.6	88.9
SVM	Normal	98.1	98.9
	DOS	97.8	98.6
	Probe	90.7	91.3
	U2R	93.7	95.9
	R2L	91.8	93.9
CART	Normal	88.9	91.9
	DOS	82.7	89.5
	Probe	82.1	85.4
	U2R	73.1	80.7
	R2L	80.8	89.0
Navie Bayes	Normal	70.3	75.9
	DOS	72.7	75.0
	Probe	70.9	75.1
	U2R	70.7	74.3
	R2L	69.8	71.1

Table II shows the test accuracy on class Normal attack that compared with 41 features and with the reduced set of features by using CFS subset technique. Here the Random Forest algorithm shows the highest accuracy compared with rest of the algorithms by considering with and without feature reduction.

VI. Conclusion and Future Work

In this paper, we have analyzed the NSL-KDD dataset that solves some of the issues of KDD cup99 data. The analysis shows that NSL-KDD dataset is very ideal for comparing different intrusion detection models. Using all the 41 features in the dataset to evaluate the intrusive patterns may leads to time consuming and it also reduce performance degradation of the system. Some of the features in the dataset are redundant and irrelevant for the process. CFS Subset is used to reduce the dimensionality of the dataset. The experiment has been carried out with different classification algorithms for the dataset with and without feature reduction and it's clear that Random Forest shows a high test accuracy compared to all other algorithms in both the cases. So in the case of reduced feature set this analysis shows that Random Forest is speeding up the training and the testing methods for intrusion detection that is very essential for the network application with a high speed and even providing utmost testing accuracy. In future we can try to improve the Random Forest algorithm to build an efficient intrusion detection system.

Reference

1. C. E. Landwehr, A. R. Bull, J. P. McDermott, and W. S. Choi, "A taxonomy of computer program security flaws," ACM Comput. Surv., vol. 26, no. 3, pp. 211–254, 1994.

2. M. Shyu, S. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM03), pp. 172–179, 2003.

3. KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup 99/kddcup99.html, October 2007.

4. J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory," ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262–294, 2000.

5. Vipin Kumar, Himadri Chauhan, Dheeraj Panwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4, September 2013.

6. "Nsl-kdd data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/KDD/NSL-KDD.html, March 2009.

7. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set", In the Proc. Of the IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), pp. 1-6, 2009.

8. Shilpa lakhina, Sini Joseph and Bhupendra verma, "Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD", International Journal of Engineering Science and Technology, Vol. 2(6), 2010, 1790-1799.

9. Lei Li, De-Zhang Yang, Fang-Cheng Shen, "A Novel Rule-based Intrusion detection System Using Data Mining", In the Proc. Of 3rd IEEE International Conference on Computer Sceince and Information Technology, pp. 169-172, 2010.

10. Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, " Top Ten Data Mining Algorithms", Knowledge and Information Systems Journal, Springer-Verlag London, vol. 14, Issue 1, pp. 1-37, 2007.

11. Weka – Data Mining Machine Learning Software.

12. http://www.cs.waikato.ac.nz/ml/weka/