Intrusion Detection System Based On Feature Removal Technique

[1] G. Vinodhini,
[1]PG Scholar, Department of CSE,
Avinashilingam Institute for Home Science and Higher Education for Women Coimbatore, TamilNadu, India.

[2]M. Dharani, [3]T. Menaka
[2,3] PG Scholar, Department of CSE,
Avinashilingam Institute for Home Science and Higher Education for Women Coimbatore, TamilNadu, India.

***Abstract*-An intrusion detection system (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a Management Station. The data features are important in determining the efficiency of the intrusion detection system. This paper proposes a feature removal method for selecting the best possible features. The performance of the Feature Removal Method is determined using the Support Vector Machine (SVM). The KDD cup 99 is used as a benchmark dataset to detect the intrusion.**

Keywords: Feature Removal Method, KDD Cup 99, Intrusion Detection, Support Vector Machine.

I. INTRODUCTION

The Internet has become a part of daily life and an essential tool today. It assists people in many areas, such as business, entertainment, education and research etc. In particular, Internet has been used as a significant part in business models. For the business, both business people and customers apply the Internet application such as website, e-mail and financial transactions on business activities. Therefore, the use of Internet and its information security has to be carefully concerned. Intrusion detection evolved as one major research problem for business and personal networks.

As there are numerous risks of network attacks in the Internet environment, there are several systems designed to block the Internet-based attacks. Particularly, intrusion detection system aids the network to resist external attacks. To overcome this problem, various artificial intelligence and machine learning methods are developed, such as fuzzy logic, K-nearest neighbor, support vector machine (SVM), artificial neural networks (ANN), Naïve Bayes networks, principal component analysis (PCA), decision tree and genetic algorithm (GA).

Among the various methods mentioned above, an effective method is SVM, which is a well-known classifier tool based on small sample learning. Since SVM has evidenced its robustness and efficiency in the network action classification, it therefore becomes a popular method widely used in IDS.

In general, IDS deals with very large amount of data which contain redundant and irrelevant features, by reducing the features training and predicting time could be saved. Methods for feature selection are divided into two categories: filter method and wrapper method. Filter methods are independent of an inductive algorithm. Whereas, Wrapper method is kind of feature removal method use a predictive model to score feature subsets. Each subset is used to train a model, which is tested on a hold-out set. By calculating the number of mistakes made on that hold-out set gives the score for that subset. Meanwhile, filter methods use measures that are independent of the predetermined classification algorithms to estimate the goodness of candidate subsets. Generally, wrapper methods generally perform better than filter methods. This paper proposes wrapper-based method in selecting vital features which is an improved wrapper based feature removal method.

This paper is organized as follows: Section 2 discusses some related work; Section 3 presents the dataset description; Section 4 provides the overview of the proposed framework; Section 5 provides the methodology; Section 6 describes our experimental result and discussion; Section 7 provides the conclusions of our work.

II. RELATED WORK

Shai rubin et al., [1] proposed a technique called promatching that combines protocol analysis, normalization and pattern matching to exclude the non-attack traffic quickly to consume the time. Marco Cova et al., [2] proposed a method called Swaddler for detection anomaly based attacks in the web application. This method analyse the internal state of a web application and learnsthe relationship between application's critical point and its internal state. Pavel kachurka et al., [3] proposed a technique to detect the intrusion in real time network data and recognise the intrusion.

Jose M. Cadenas et al., [4] proposed approach is based on a Fuzzy Random Forest and it integrates filter and wrapper methods into a sequential search procedure with improved classification accuracy of the features selected. Md. MonirulKabir et al., [5] proposed a new feature selection (FS) algorithm based on the wrapper approach using neural networks (NNs). The vital aspect of this algorithm is the automatic determination of NN architectures during the FS process. This algorithm uses a constructive approach involving correlation information in selecting features and determining NN architectures. Ron

Kohavi, et al., [6] described the feature subset selection problem in supervised learning, which involves identifying the relevant or useful features in a dataset and giving only that subset to the learning algorithm.

Edgar Osuna et al., [7] proposed a decomposition algorithm that is guaranteed to solve the QP problem and that does not make assumptions on the expected of support vectors. They considered a foreign exchange rate time series data base with 110,000 data points that generate 100,000 support vectors to present the feasibility of their approach. Thorsten Joachims et al., [8] discussed about the improved algorithm for training SVM on large learning tasks with many training examples off-the-shelf optimization techniques for general quadratic programs quickly become intractable in their memory and time requirements. SVM light is an implementation of an SVM learner which addresses the problem of large tasks to overcome this they presented a algorithmic and computational results developed for SVMlightV2.0. The results give guidelines for the application of SVMs to large domains.

Theodoros Evgeniou et al., [9] discussed that the Regularization Networks and Support Vector Machines are techniques for solving certain problems of learning from examples – in particular the regression problem of approximating a multivariate function from sparse data. They reviewed both regularization and Support Vector Machines formulations in the context of Vapnik's theory of statistical learning that provides a general foundation for the statistics, learning problem and combining functional analysis. Hiep-Thuan Do et al., [10] proposed a new incremental, parallel and distributed SVM algorithm using linear or non-linear kernels which aims at classifying very large datasets on standard personal computers they extended the recent finite Newton classifier for building an incremental, parallel and distributed SVM algorithm. This algorithm is very fast and can handle very large datasets in linear or nonlinear classification tasks.

Srinivas ,M et al., [11] reviewed that SVM are learning machines that plot the training vectors in high dimensional feature space, labelling each vector by its class. SVM method classifies data by determining a set of support vectors, which are members of training inputs. Computing the hyper plane to separate the data points leads to a quadratic optimization problem. There are two main reasons that we used SVMs for intrusion detection. They are its performance in terms of execution speed, and its scalability. SVMs are relatively insensitive to the number of data points and the classification complexity does not depend on the dimensionality of the feature space.

III. DATASET DESCRIPTION

The KDD99 dataset is most commonly used in Knowledge Discovery and Data Mining Tools Competition for building network intrusion detector and data between intrusions and normal network connections is differentiated [12]. This KDDcup99 dataset is derived from DARPA intrusion detection evaluation program. For acquiring raw TCP/IP dumps data for a local-area network (LAN), the simulation environment was set up by the MIT Lincoln Lab. The KDD99 dataset contains both labelled and unlabelled records. The labelled record consists of 41 attributes.

In KDD99 dataset, each example represents attribute values of a class in the network data flow, and each class is named either normal or attack. The classes in KDD99 dataset classified into five types (normal, probe, DOS, U2R, and R2L).

- Denial of Service Attack (DoS): is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legal requests, or denies legal users access to a machine.
- User to Root Attack (U2R): is a class of exploit in which the attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to gain root access to the system.
- Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.
- Probing Attack: is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

A complete KDD'99 dataset comprises five millions connection records where 4,898,431 are labelled connections that divided into 22 different attack classes that are tabulated in Table 1.

Table I Details of Attacks of Labelled Records

Category of Attack	Attack Name
Normal	Normal
DoS	Neptune,Smurf,Pod,Teardrop,Land,back
Probe	Portsweep, IPsweep, Nmap, satan
U2R	Bufferoverflow,LoadModule,Perl,Rootkit
R2L	Guesspassword,Ftpwrite,Imap,Phf, Multihop,Warezmaster,Warezclient

IV. OVERVIEW OF THE PROPOSED FRAMEWORK

The proposed framework based on analysis of KDD cup 99 dataset. Even after Pre-processing the data there is a difficulty in dimensionality problem which leads to high processing time and low efficiency. To overcome these problems this paper proposed a method in reducing the dimension of data feature. By using the Feature Removal method the best feature are selected and evaluated using the SVM classifier. Figure 1 shows the proposed framework.

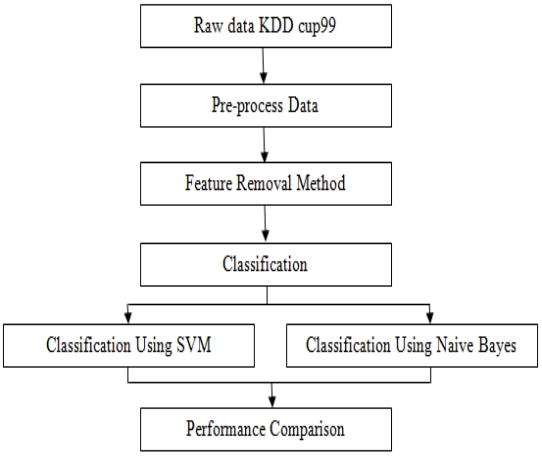

Fig. 1: Proposed System Architecture

V. METHODOLOGY

In this paper we proposed a method called Feature removal method based on wrapper method. Pre-processed KDD cup99 dataset maps to a mathematical vector with 41 features. The pre-processed data are streamlined into machine learning method**.** In the first iteration all the features are evaluated using the classifier, then by deleting one feature, update the dataset and using the classifier efficiency, the importance of the provided feature is calculated. In the second iteration the second feature is deleted while the first feature is kept as it is in the dataset. The process is repeated until all 41 features are evaluated. After calculating the entire feature's efficiency, it is sorted and the vital feature is selected. The efficiency of the selected features is compared with the total feature efficiency.

Algorithm: Feature Removal Method

Input: Dataset X with n Features
Output: Vital features
BEGIN
1. Let X = $\{x_1, x_2, \ldots, x_n\}$,where n represents the features of the dataset.
2. FOR (i=1,2,...,n), do
Delete x_i
Update $X^{(i)} = (x_1, x2 \ldots xi_{-1}, xi_{+1}, \ldots, x_n)$ as new feature
Apply Classification Algorithm
End FOR
3. Sort and select vital features based on classifier accuracy
END

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental evaluation is demonstrated on KDD cup99 training dataset to verify the effectiveness and performance of the proposed system. The dataset contains both the data of normal and abnormal class. The dataset is tested with trained SVM. The classifier is evaluated with 10-fold cross validation for estimating the performance. It is found that the accuracy of the selected feature is higher than total feature accuracy and the execution time and speed of the system is less when compared with total feature data. The result of Feature removal method using SVM classifier is shown in figure 2. The accuracy of the DoS attack after reducing the feature is increased up to 0.05% by using only best 12 attributes. In the same way the accuracy of probe attack is increased up to 0.28% after selecting the best 12 features. The accuracy of R2L attack and U2R attack remains the same. These two R2L and U2R attack uses only 12 features to attain the result. Henceforth the result of Feature Removal Method using Naive Bayes classifier is shown in figure 3. After applying feature removal method only best 12 features are used for evaluating the accuracy. The accuracy of DoS attack is increased to 0.15% after selecting the best features using the Feature Removal Method. The accuracy of probe attack is increased up to 3.5%. Similarly the accuracy of R2L attack after reducing the feature from whole feature set is increased up to 0.9%. The accuracy of U2R attack drastically increased up to 5.25%. The experimental result shows that the accuracy of detecting the attacks achieved is greater than that of total feature accuracy. It is clear that the comparison result of the SVM classifier and Naive Bayes classifier shown in figure 4 that the performance of SVM on reduced dataset is better than Naive bayes method. The experimental results show that the proposed system is faster and efficient.

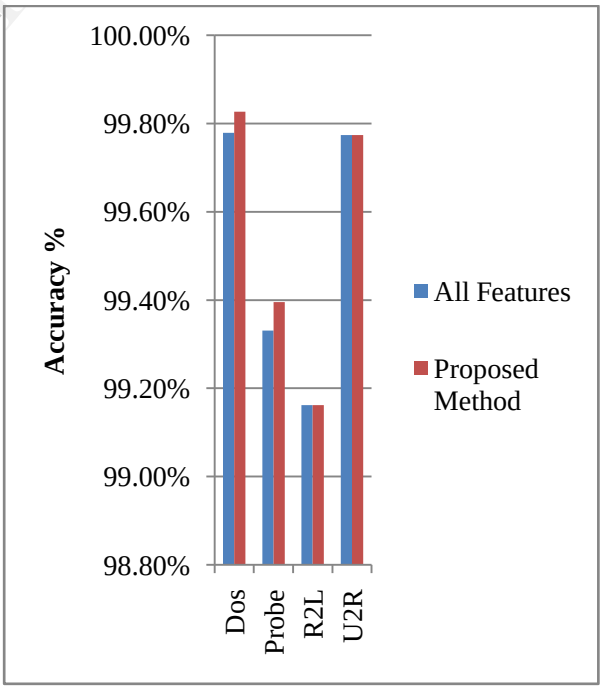

Fig. 2: Performance of the Algorithm Using SVM Classifier

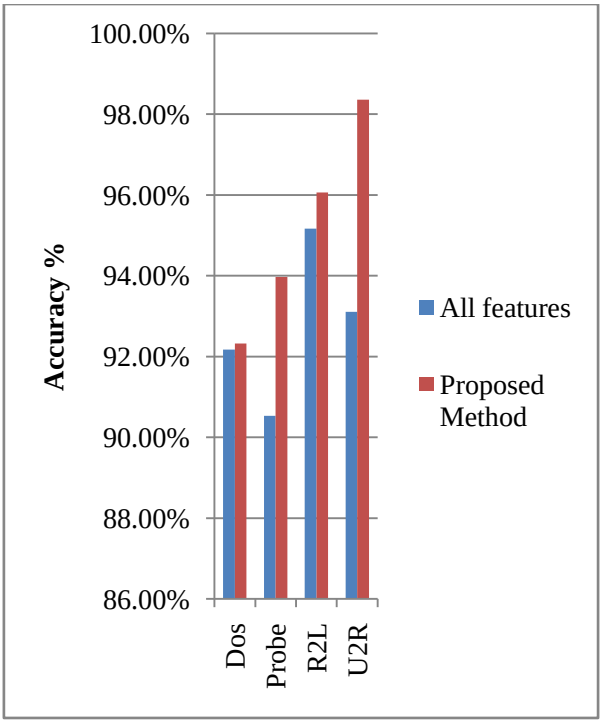

Fig. 3: Performance of the Algorithm Using Naive Bayes Classifier

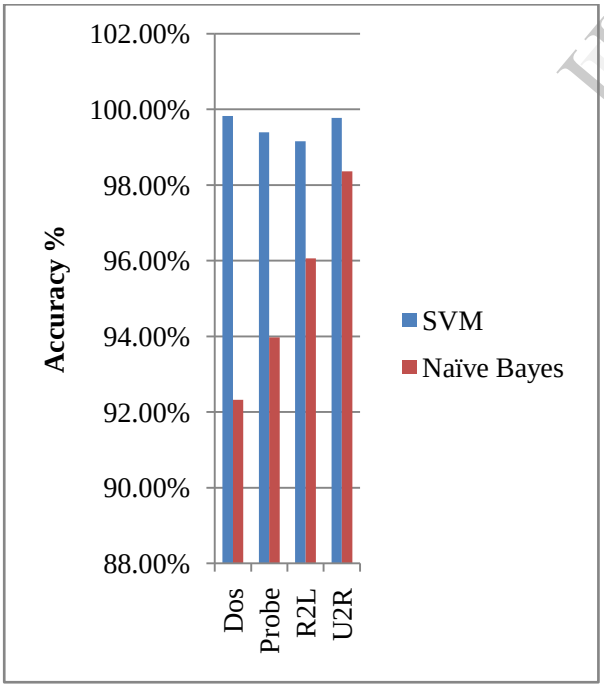

Fig. 4: Accuracy Comparison of the Algorithm using SVM and Naive Bayes Classifier

VII. CONCLUSION

This paper statistically analyzed the performance of complete KDD dataset and proposed a method to reduce the dimensionality of the feature. Since the data features are important in determining the efficiency of the intrusion detection system it is important to reduce the feature to improve the performance of the system. The Feature Removal Method reduced the attributes of the dataset by evaluating the importance of each feature. The reduced dataset is evaluated using the SVM and Naive Bayes classifier. The experimental result shows that the performance of the proposed method using the Support Vector Machine than outperformed the other Classification techniques.

ACKNOWLEDGEMENT

Vinodhini G received her B.E in Information Technology from Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India in 2010 and is currently pursuing her M.E in Computer Science and Engineering in Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.

Dharani M received her B.E in Computer Science and Engineering from Kalasalingam University, Srivilliputhur, India in 2012 and is currently pursuing her M.E in Computer Science and Engineering in Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.

Menaka T received her B.E in Computer Science and Engineering from Vivekananda College of Engineering, Thiruchengode, India in 2012 and is currently pursuing her M.E in Computer Science and Engineering in Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.

REFERENCE

[1] Shai rubin, Somesh jha and Barton P. Miller, "Promatching Network Traffic for High Throughput Network Intrusion Detection", In proceedings of the 13th ACM conference on computer and communications security, pages 47-58, 2006.

[2] Marco cova, Davide Balzarotti, viktoria felmetsger,and Giovanni vigna Swaddler, " An Approach for the Anomaly Based Detection," Symposium on Recent Advances in Intrusion Detection(RAID), pages 63-86. Springer, 2007.

[3] Pavel kachurka, Vladimir Golovko, "Neural Network Approach to Real-Time Network Intrusion Detection and Recognition," The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing System (IDAACS): Technology and Application, 15-17 September 2011, pages.393-397.

[4] Jose M. Cadenas, M. Carmen Garrido, Raquel Martínez "Feature subset selection Filter–Wrapper based on low quality data", Expert Systems with Applications 40, pages 6241–6252, 2013.

[5] Md. MonirulKabir, Md.MonirulIslam, KazuyukiMurase "A new wrapper feature selection approach using neural network", Neurocomputing 73, pages 3273–3283, 2010.

[6] Ron Kohavi , George H. John "Wrappers for feature subset selection", Artificial Intelligence , pages 273-324, B.V 1997.

[7] Edgar Osuna, Robert Freund and Federico Girosi , "An Improved Training algorithm for support Vector Machines", Proceedings of IEEE workshop on Neural Network Signal Processing, amelia Island, FL 24—26 September 1997.

[8] Thorsten Joachims, "Making Large-Scale SVM Learning Practical", Advances in Kernel Methods - Support Vector Learning, Artificial Intelligence –Unit, MIT Press, Cambridge, USA, pages 169-184, 1998.
[9] Theodoros Evgeniou, Massimiliano Pontil and Tomaso Poggio, "Regularization Networks and Support Vector Machines", Advances in Computational Mathematics, Volume 13, Issue 1, pages 1-50 2000.
[10] Hiep-Thuan Do, Nguyen-Khang Pham and Thanh-Nghi Do, "A Simple, Fast Support Vector Machine Algorithm for Data Mining", Fundamental & Applied IT Research Symposium 2005.
[11] Srinivas ,M., Guadalupe,J., and Amdrew,.S. "Intrusion Detection using Neural Networks and Support Vector Machines". In Proceedings of the International Joint Conference on Neural Networks, 2002.
[12] MIT Lincoln Labs, 1998 DARPA Intrusion Detection Evaluation. Available on: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/ index.html, February 2008.