International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

New Approach for Intrusion Detection Using Naive Bayes Filter-Based Feature Selection Algorithm

Priyanka Bhoite
Department of Computer Engineering, Zeal College of Engineering and Research, Pune. Savitribai Phule Pune University

Abstract: ***Repetitive and unessential components in information have brought about a long haul issue in system activity grouping. These components back off the procedure of arrangement as well as keep a classifier from settling on precise choices, particularly when adapting to enormous information. In this paper, we propose a shared data based calculation that systematically chooses the ideal component for arrangement. This shared data based component determination calculation can deal with directly and non linearly subordinate information highlights. Its adequacy is assessed in the instances of system interruption discovery. An Intrusion Detection System (IDS), named Least Square Support Vector Machine based IDS(LSSVM-IDS), is fabricated utilizing the elements chose by our proposed include determination calculation. The execution of LSSVM-IDS is assessed utilizing three interruption identification assessment datasets, to be specific KDD Cup 99, NSL-KDD and Kyoto 2006+dataset. The assessment comes about demonstrate that our element choice calculation contributes more basic elements for LSSVM-IDS to accomplish better precision and lower computational cost contrasted and the best in class techniques. The contribution work is, replacing the classification algorithm. In existing system, we are using LS-SVM classifiers for classification but as compare to naive Bayes classifier and LS-SVM classifier naive Bayes classifier, naive Bayes classifier algorithm accuracy of classification is high and it is to identify important reduced input features in building IDS that is computationally efficient and effective.***

Keywords: Feature Selection, Intrusion Detection, Least Square Support Vector Machine, Linear Correlation Coefficient, Mutual Information, Naive Bayes Algorithm.

I. INTRODUCTION

Despite increasing awareness of network security, the existing solutions remain incapable of fully protecting Inter-net applications and computer networks opposite the threats from ever-advancing cyber attack method like as DoS attack and computer malware. Developing effective and adaptive security approaches, therefore, has become more critical than ever before. The traditional security techniques, as the first line of security defense, such as user authentication, firewall and data encryption, are insufficient to fully cover the hole landscape of network security while facing challenge from ever-evolving intrusion skills and method [1]. Hence, other line of security defense is more recommended, like Intrusion Detection System (IDS). Currently, an IDS alongside with anti-virus software has become an important complement to the security infrastructure of most organizations. The combination of these two lines provides a more comprehensive defense against those threats and enhances network security. A significant amount of research has been conducted to develop intelligent intrusion detection techniques, which help achieve better network security. Bagged boosting-based on C5 decision trees [2] and Kernel Miner [3] are two of the earliest attempts to build intrusion detection schemes. Methods proposed in [4] and [5] have successfully applied machine learning techniques, such as Support Vector Machine (SVM), to classify network traffic patterns that do not match normal network traffic. Both systems were equipped with five distinct classifiers to detect normal traffic and four different types of attacks (i.e., DoS, probing, U2R and R2L). Experimental results show the effectiveness and robustness of utilizing SVM in IDS. Mukkamala et al. [6] researched the possibility of assembling different learning strategies, including Artificial Neural Networks (ANN), SVMs and Multivariate Adaptive Regression Splines (MARS) to detect intrusions. They pre-pared five different classifiers to recognize the normal traffic from the four different types of attacks. They compared the performance of each of the learning strategies with their model and found that the ensemble of ANNs, SVMs and MARS accomplished the best execution in terms of classification accuracies for all the five classes. Toosi et al. [7] combined a arrangement of neuro-fuzzy classifiers in their design of a detection framework, in which a genetic algorithm was applied to optimize the structures of neuro-fuzzy framework utilized in the classifiers. Based on the pre-determined fuzzy inference framework (i.e., classifiers), detection choice was made on the incoming traffic. Recently, we proposed an anomaly-based scheme for detecting DoS attacks [8]. The system has been evaluated on KDD Cup 99 and ISCX 2012 datasets and achieved promising identification accuracy of 99.95% and 90.12% respectively.

However, current network traffic data, which are often huge in size, present a major challenge to IDSs [9]. These big data slow down the entire detection process and may lead to unsatisfactory classification accuracy due to the computational difficulties in handling such data. Classifying a huge amount of data usually causes many mathematical difficulties which then lead to higher computational complexity. As a well-known intrusion calculation dataset, KDD Cup 99 dataset is a typical example of more-scale datasets. This dataset contains of more than five million of training samples and two million of testing samples respectively. Such a large scale dataset check the building and testing procedure of a classifier, or form the classifier unable to do due to framework failures caused by low memory. Furthermore, large-scale datasets usually contain noisy, redundant, or uninformative features which present critical challenges to knowledge discovery and information modeling.

II. LITERATURE SURVEY

Feature selection is a technique for eliminating irrelevant and redundant features and selecting the most optimal subset of features that produce a better characterization of patterns belonging to different classes. Methods for feature selection are generally classified into filter and wrapper methods [2].

Filter algorithms utilize an independent measure (such as, information measures, distance measures, or consistency measures) as a criterion for estimating the relation of a set of features, while wrapper algorithms make use of particular learning algorithms to evaluate the value of features. In comparison with filter methods, wrapper methods are often much more computationally expensive when dealing with high-dimensional data or large-scale data. In this study hence, we focus on filter methods for IDS. Due to the continuous growth of data dimensionality, feature selection as a pre-processing step is becoming an essential part in building intrusion detection systems [3].

Mukkamala and Sung [4] proposed a novel feature selection algorithm to reduce the feature space of KDD Cup 99 dataset from 41 dimensions to 6 dimensions and evaluated the 6 selected features using an IDS based on SVM. The results show that the classification accuracy increases by 1% when using the selected features. Chebrolu et al. [5] inspected the performance in the utilize of a Markov blanket model and decision tree analysis for feature selection, which present its capability of decreasing the number of features in KDDCup 99 from 41 to 12 features.

Chen et al. [6] present an IDS based on Flexible Neural Tree (FNT). The model applied a pre-processing feature selection phase to increase the detection performance. utilizing the KDD Cup 99, FNT model achieved 99.19% detection correctness with only 4 features. Recently, Amiri [2] present a forward feature choice algorithm utilizing the mutual data strategy to calculate the relation between features. The optimal feature set was then utilize to train the LS-SVM classifier and make the IDS. Horng et al. [7] proposed an SVM-based IDS, which combines a hierarchical clustering and the SVM. The hierar-chical clustering algorithm was used to provide the classifier with fewer and higher quality training data to reduce the average training and testing time and improve the classification performance of the classifier. Experimented on the corrected labels KDD Cup 99 dataset, which includes some new attacks, the SVM-based IDS scored an overall accuracy of 95.75% with a false positive rate of 0.7%.

All of the aforementioned detection techniques were eval-uated on the KDD Cup 99 dataset. However, due to some limitations in this dataset, which will be discussed in Subsec-tion 5.1, some other detection methods [8], [9] were evaluated using other intrusion detection datasets, such as NSL-KDD and Kyoto 2006+. A dimensionality reduction method proposed in [11] was to find the most important features involved in building a naive Bayesian classifier for intrusion detection. Experiments conducted on the NSL-KDD dataset produced encouraging results. Chitrakar et al. [10] proposed a Candidate Support Vector based Incremental SVM algorithm (CSV-ISVM in short). The algorithm was applied to network in-trusion detection. They evaluated their CSV-ISVM-based IDS on the Kyoto 2006+ [11] dataset. Experimental results showed that their IDS produced promising results in terms of detection rate and false alarm rate. The IDS was claimed to perform realtime network intrusion detection. Therefore, in this work, to make a fair comparison with those detection systems, we evaluate our proposed model on the aforementioned datasets.

III. PROPOSED SYSTEM

The framework of the proposed intrusion detection system is depicted in Figure 1. The detection framework is comprised of four main phases:

Data collection, where sequences of network packets are collected,

Data prepossessing, where training and test data are preprocessed and important features that can distinguish one class from the others are selected.

Classifier training, where the model for classification is trained using LS-SVM.

Attack recognition, where the trained classifier is used to detect intrusions on the test data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

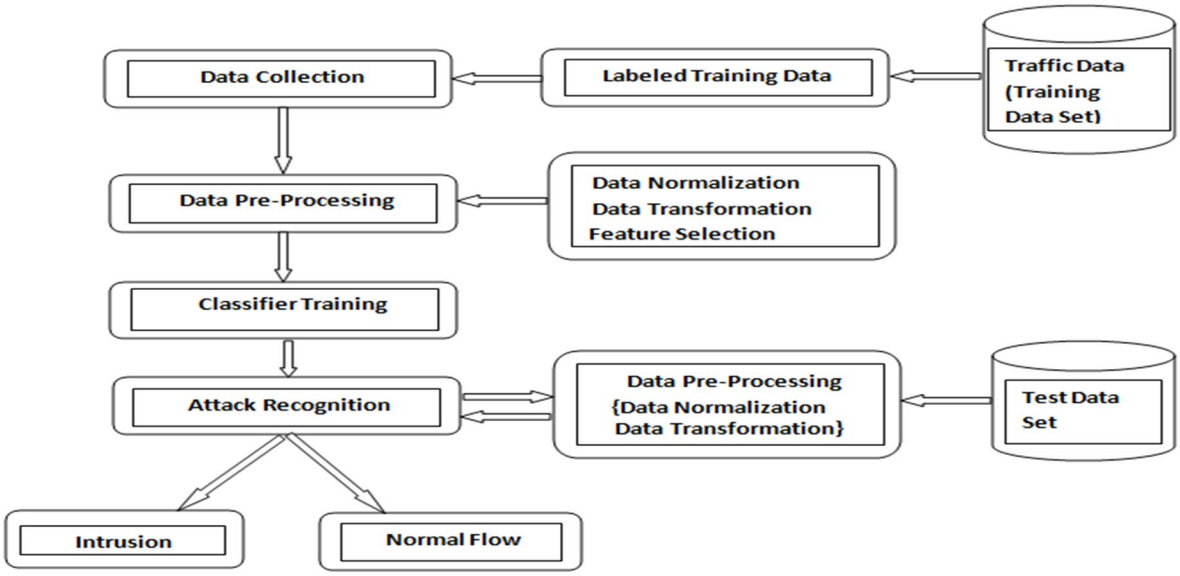

Fig. 1. System Architecture

A. Data Collection

Data collection is the first and a critical step to intrusion detection. The type of data source and the location where data is collected from are two determinate factors in the design and the effectiveness of an IDS. To provide the best suited protection for the targeted host or networks, this study proposes network-based IDS to test our proposed approaches. The proposed IDS execute on the neighbourhood router to the victim(s) and monitor the inbound network traffic. During the training stage, the collected information samples are distinguish with respect to the transport/Internet layer protocols and are labeled opposite the domain knowledge. However, the information collected in the test stage are categorized like to the protocol types only.

B. Data Preprocessesing

The data obtained during the phase of data collection are first processed to generate the basic features such as the ones in KDD Cup 99 dataset. This phase contains three main stages shown as follows.

1) *Data Transferring:* The trained classifier requires each record in the input data to be represented as a vector of real number. Thus, every symbolic feature in a dataset is first converted into a numerical value. For example, the KDD CUP 99 dataset contains numerical as well as symbolic features. These symbolic features include the type of protocol (i.e., TCP, UDP and ICMP), service type (e.g., HTTP, FTP, Telnet and so on) and TCP status flag (e.g., SF, REJ and so on). The method simply replaces the values of the categorical attributes with numeric values.
2) *Data Normalization:* An essential step of data prepossessing after transferring all symbolic attributes into numerical values is normalization. Data normalization is a process of scaling the value of each attribute into a well-proportioned range, so that the bias in favor of features with greater values is eliminated from the dataset. Data used in Section 5 are standardized. Every feature within each record is normalized by the respective maximum value and falls into the same range of [0-1]. The transferring and normalization process will also be applied to test data. For KDD Cup 99 and to make a comparison with those systems that have been evaluated on different types of attacks, we construct five classes. One of these classes contains purely the normal records and the other four hold different types of attacks (i.e., DoS, Probe, U2R, R2L), respectively.
3) *Feature Selection:* Even though every connection in a dataset is represented by various features, not all of these features are needed to build an IDS. Therefore, it is important to identify the most informative features of traffic data to achieve higher performance. In the previous section using Algorithm 1, a flexible method for the problem of feature selection, FMIFS, is developed. However, the pro-posed feature selection algorithms can only rank features in terms of their relevance but they cannot reveal the best number of features that are needed to train a classifier. Therefore, this study applies the same technique proposed in to determine the optimal number of required features. To do so, the technique first utilizes the proposed feature selection algorithm to rank all features based on their importance to the classification processes. Then, incrementally the technique adds features to the classifier one by one. The final decision of the optimal number of features in each method is

taken once the highest classification accuracy in the training dataset is achieved. The selected features for all datasets, where each row lists the number and the indexes of the selected features with respect to the corresponding feature selection algorithm. In addition, for KDD Cup 99, the proposed feature se-lection algorithm is applied for the aforementioned classes.

4) *Classifier Training:* Once the optimal subset of features is selected, this subset is then taken into the classifier training phase where LS-SVM is employed. Since SVMs can only handle binary classification problems and because for KDD Cup 99 five optimal feature subsets are selected for all classes, five LS-SVM classifiers need to be employed. Each classifier distinguishes one class of records from the others. For example the classifier of Normal class distinguishes Normal data from non-Normal (All types of attacks). The DoS class distinguishes DoS traffic from non-DoS data (including Normal, Probe, R2L and U2R instances) and so on. The five LS-SVM classifiers are then combined to build the intrusion detection model to distinguish all different classes.

5) *Attack Recognition:* In general, it is simpler to build a classifier to distinguish between two classes than considering multi classes in a problem. This is because the decision boundaries in the first case can be simpler. The first part of the experiments in this paper uses two classes, where records matching to the normal class are reported as normal data, otherwise are considered as attacks.

IV. ALGORITHMS

A. *Proposed System Module*

1) *Admin:* There are 5 steps to detect intrusion using new approach algorithm.
2) First collecting the training labelled data from training dataset.
3) After, process the collected data, in pre-processing for feature extraction using flexible mutual information based feature selection or flexible linear correlation coefficient based feature selection algorithm.

a) Data Transformation
b) Data Normalization
c) Feature Selection
d) After feature selection, classify the data using LS-SVM Algorithm and Naive Bayes algorithm.
e) After classification, recognize the attack using Attack classification based on LS-SVM algorithm and Naive Bayes algorithm.
f) After, displaying results for attacks list.

B. *Proposed System Algorithm*

Step 1: First we collect the training labeled data from training dataset.
Step 2: Second we process the collected data, In pre-processing for feature extraction we use flexible mutual information based feature selection or flexible linear correlation coefficient based feature selection algorithm
Step 3: After feature selection we classify the data using LS-SVL Algorithm.
Step 4: After classification recognize the attack using Attack classification based on LS-SVM algorithm.Step 5: Finally results demonstrate that the intrusion detected or not.Step 6: Stop.

V. PERFORMANCE MEASURE

The classification performance of the intrusion detection model combined with FMIFS, MIFS (= 0:3), MIFS (= 1) and FLCFS and the model using all features based on the three datasets shown in existing work. The results clearly demonstrate that the classification performance of an IDS is enhanced by the feature selection step. In addition, the proposed feature selection algorithm FMIFS shows promis-ing results in terms of low computational cost and high classification results. It determine clearly that the searching model combined with the FMIFS has give an accuracy rate of 99.79%, 99.91% and 99.77% for KDD Cup 99, NSL-KDD and Kyoto 2006+, respectively, and significantly detection model combined with FMIFS enjoys the highest detection rate and the lowest false positive rate in comparison with other combined detection models. The proposed feature choice algorithm is computationally capable when it is put to the LSSVM-IDS. the LSSVM-IDS + FMIFS performs good than LSSVM-IDS with all 41 features on all datasets. There are significant differences when performing experiments on KDD Cup 99 and NSL-KDD and a slight difference on Kyoto 2006+ dataset by comparison with the two aforementioned models. Here, DR denotes detection rate, FPR denotes false positive rate. Fig.2 displays performance of building and testing times of LSSVM-IDS using all features and LSSVM-IDS combined with FMIFS.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

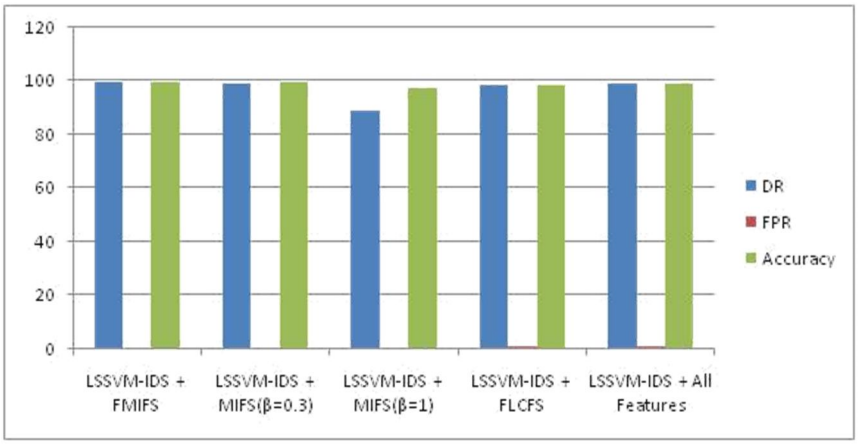

Fig. 2. Building and testing times of LSSVM-IDS using all features and LSSVM-IDS combined with FMIFS

B. Result Table

Table : Performance classification for all attacks based on the dataset

System	DR	FPR	Accuracy
LSSVM-IDS + FMIFS	99.46	1.25	99.79
LSSVM-IDS + MIFS(=0.3)	99.38	1.23	99.70
LSSVM-IDS + MIFS(=1)	89.26	1.34	97.63
LSSVM-IDS + FLCFS	98.47	1.61	98.41
LSSVM-IDS + All Features	99.16	1.97	99.19

VI. SYSTEM ANALYSIS

The proposed feature selection algorithm is computationally efficient when it is applied to the LSSVM-IDS. Figure2 shows the building (training) and test time consumed by the detection model using FMIFS compared with the detection model using all features. The figure shows that the LSSVM-IDS + FMIFS performs better than LSSVM-IDS with all features on dataset. In this paper, a supervised filter-based feature selection algorithm has been proposed, namely Flexible Mutual Information Feature Selection (FMIFS). FMIFS is an improvement over MIFS and MMIFS. FMIFS suggests a modification to Battitis algorithm to reduce the redundancy among features.

VII. CONCLUSION

In this paper, we are replacing the classification algorithm. In existing system, we are using LS-SVM classifiers for classification but as compare to naive Bayes classifier and LS-SVM classifier, naive Bayes classifier algorithm accuracy is high and it is important to reduced input features in building IDS that is computationally efficient and effective. Efficiency is medium for LS-SVM classifier but high in naive Bayes classifier. SVM is useful when your classification categories are fixed, but naive Bayes classifier is useful when your classification categories are variable or dynamically changed. There are two main components are essential to build an IDS. They are a robust classification method and an efficient feature selection algorithm. The proposed system, supervised filter-based feature selection algorithm has been implemented, namely Flexible Mutual Information Feature Selection (FMIFS). FMIFS is an improvement over MIFS and MMIFS. FMIFS suggests a modification to Battitis algorithm to reduce the redundancy among features. FMIFS eliminates the redundancy parameter required in MIFS and MMIFS. Finally, based on the experimental results achieved on dataset, it can be concluded that the proposed detection system has achieved promising performance in detecting intrusions over computer networks.

REFERENCES

[1] R. Battiti, "Using mutual data for selecting features in supervised neural net learning", IEEE Transactions on Neural Networks 5 (4) (1994) 537550.
[2] F. Amiri, M. Rezaei Yousefi, C. Lucas, A. Shakery, N. Yazdani," Mutual data-based feature selection for intrusion detection method", Journal of Network and Computer Applications 34 (4) (2011) 11841199.
[3] A. Abraham, R. Jain, J. Thomas, S. Y. Han," D-scids: Distributed soft computing intrusion detection method", Journal of Network and Computer Applications 30 (1) (2007) 8198
[4] S. Mukkamala, A. H. Sung, "Significant feature selection utilizing computational intelligent strategy for intrusion detection, in: Advanced Methods for Knowledge Discovery from Complex Data", Springer, 2005, pp. 285306.
[5] S. Chebrolu, A. Abraham, J. P. Thomas," Feature deduction and ensemble design of intrusion detection systems", Computers Security 24 (4) (2005) 295307.
[6] Y. Chen, A. Abraham, B. Yang, "Feature selection and classification flexible neural tree", Neurocomputing 70 (1) (2006) 305313. [7] S.-J. Horng, M.-Y. Su, Y.-H. Chen, T.-W. Kao, R.-J. Chen, J.-L. Lai, C. D. Perkasa, A novel intrusion detection system based on hierarchical clustering and support vector machines, Expert systems with Applications 38 (1) (2011) 306313.
[7] G. Kim, S. Lee, S. Kim," A novel hybrid intrusion detection method integrating anomaly detection with misuse detection", Expert Systems with Applications 41 (4) (2014) 16901700.
[8] P. Gogoi, M. H. Bhuyan, D. Bhattacharyya, J. K. Kalita, "Packet and flow based network intrusion dataset", in: Contemporary Computing, Vol. 306, Springer, 2012, pp. 322334.
[9] R. Chitrakar, C. Huang," Selection of candidate support vectors in incremental svm for network intrusion detection", Computers Security 45 (2014) 231241.
[10] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, K. Nakao, "Statistical analysis of honeypot data and building of kyoto 2006+ dataset for nids evaluation, in: Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security", ACM, 2011, pp. 2936.