International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

A Proposed Approach for Network Intrusion Detection System using SVM

Miss Pooja Deshmukh and Dr. S. B. Javheri

Department of Computer Engineering JSPM's Rajarshi Shahu College Of Engineering Savitribai Phule Pune University Pune, India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

A system Network Intrusion Discovery Framework (NIDS) helps the system admin to identify network security breaks in their own association. Nonetheless, numerous difficulties emerge while building up an intelligent and powerful NIDS for unexpected and capricious attacks. In recent years, one of the foremost focuses inside NIDS studies has been the application of machine learning knowledge of techniques. 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 nonlinearly 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 JRipper Intrusion Detection System (JRIP-IDS), is fabricated utilizing the elements chose by our proposed include determination calculation. The execution of JRIP-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 JRIP-IDS to accomplish better precision and lower computational cost contrasted and the best in class techniques.

Keywords: *Intrusion detection, Feature selection, Linear correlation coefficient, Least square support vector machine.*

Introduction

The existing solutions over network security problems of protecting computer network security are remain intractable despite of enhancement in awareness of network security. In 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 way of security defense is more recommended, like Intrusion Detection System (IDS). Currently, an IDS alongside with anti-virus software has become a crucial element towards the security infrastructure of most esteemed organizations. It is the ultimate way to provide 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 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 of utilizing Support Vector Machine (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 prepared 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 an arrangement of neuro-fuzzy classifiers in their design of a detection framework, in which a genetic algorithm was applied to optimize the structures of neurofuzzy 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, the

paper proposed an anomaly-based scheme for detecting DoS attacks [8]. The system has been evaluated on KDDCup 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][10]. 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 wellknown 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[11]. 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 modelling.

Review of Literature

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 preprocessing 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] investigated the performance in the use of a Markov blanket model and decision tree analysis for feature selection, which showed its capability of reducing the number of features in KDD Cup 99 from 41 to 12 features. Chen et al. [6] proposed an IDS based on Flexible Neural Tree (FNT). The model applied a pre-processing feature selection phase to improve the detection performance. Using the KDD Cup 99, FNT model achieved 99.19% detection accuracy with only 4 features.Recently, Amiri [2] proposed a forward feature selection algorithm using the mutual information method to measure the relation among features. The optimal feature set was then used to train the LS-SVM classifier and build the IDS.Horng et al. [7] proposed an SVM-based IDS, which combines a hierarchical clustering and the SVM. The hierarchical 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 evaluated on the KDD Cup 99 dataset. However, due to some limitations in this dataset, which will be discussed in Subsection 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 anaive 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 intrusion detection.They evaluated their CSV-ISVM-based IDS on the Kyot 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. The Detection framework buttress computer security and terminate detrimental effects on entire security running in network.

Proposed Methodology

The detection framework is mainly categorized into four phases which given below. In this module all phases are mentioned for brief clarification of computer network security. The phases are:

1) Data collection is a first and critical step to intrusion detection, where sequences of network packets are collected, 2) Data prepossessing, where training and test data are preprocessed and important features that can distinguish one class from the others are selected,
3) Classifier training, where the model for classification is trained using LS-SVM, and JRip, where optimal subset of
feature is selected,
4) Attack recognition, where the trained classifier is used to detect intrusions on the test data.

A. Advantages

1) Due to machine learning technique, it improves accuracy of intrusion detection system.
2) The network or device is continuously monitored for any invasion or attack.
3) The system may be modified and modified in step with desires of unique client and can help outside as well as inner threats to the system and network.
4) It presents user friendly interface which allows easy protection management systems.
5) Any alterations to files and directories on the machine can be easily detected and reported.
6) The system also detects certain well-known attacks and gives warnings to the user.

B. Architecture

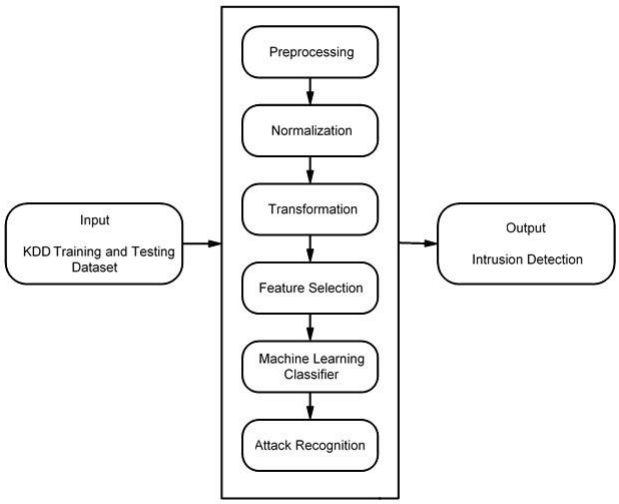

Fig. 1. Proposed System Architecture

C. Mathematical Model

Preprocessing:
In this step, training data source (T) is normalized to be equipped for processing by using following steps:
$T_{norm} = \{\frac{T-\mu_T}{\sigma_T},\ \sigma_T \neq 0\ and\ T-\mu_T,\ \sigma_T = 0$ (1)
Where,
$T = \{x_{i,j} | i = 1,2,...,m\ and\ j = 1,2,3,...,n\}$ $\mu_T = \{\mu_j | j = 1,2,3,...,n\}$ $\sigma_T = \{\sigma_j | j = 1,2,3,...,n\}$
T is m samples with n column attributes; x_{ij} is the jth column attribute in ith sample, μ_T and σ_T are $1 * n$ matrix which are the training data mean and standard deviation respectively for each of the n attributes. Test dataset (TS) which is used to measure detection accuracy is normalized using the same μ_T and σ_T as follows:
$TS_{norm} = \frac{\sigma_T}{\sigma_T}(x)/\sigma_T,\ \sigma_T \neq 0\ and\ TS-\mu_T, \sigma_T = 0$ (2)
Feature Selection:
NDAE is an auto-encoder featuring non-symmetrical multiple hidden layers. The proposed NDAE takes an input vector $x \in R^d$ and step-by-step maps it to the latent representations $h_i \in R^d$ (here d represents the dimension of the vector) using a deterministic function shown in (3) below:

$h_i = \sigma(W_i.h_{i-1} + b_i);\ i = 1,n,$ (3)

Here, $h_0 = x$, σ is an activation function (in this work use sigmoid function $\sigma(t) = 1/(1 + e^{-t})$ and n is the number of hidden layers. Unlike a conventional Auto-Encoder and Deep Auto-Encoder, the proposed NDAE does not contain a Decoder and its output vector is calculated by a similar formula to (4) as the latent representation.

$y = \sigma(Wn+1.hn + bn+1)$ (4)

The estimator of the model $\theta = (W_i, b_i)$ can be obtained by minimizing the square reconstruction error over m training samples $(x^{(i)}, y^{(i)})_{i=1}^m$, as shown in (5).
$E(\theta) = \sum_{i=1}^m (x^{(i)}, y^{(i)})^2$ (5)

D. Algorithms

1. JRIP Classifier:

JRip popularly known as Repeated Incremental Pruning to Produce Error Reduction (RIPPER) is one of the basic and most popular algorithms. In this algorithm the five attack Classes are examined in increasing size and an initial set of rules for each class is generated using incremental reduced error i.e growing of one rule by adding combination of attributes in the antecedents to the rule. Here all possible values of each attributes gets tested and then the rule is finalized. Similarly pruning step also results in dropping attributes from antecedents until the minimum possible attributes are remaining to generate the rule. The rules are selected based on information gain. The algorithm terminates on generation of rules for the five attack classes. The stategy of replacing and revising the rules hence improves the accuracy of the generated rules. The entire network intrusion detection framework is developed using WEKA environment with java packages. Once the algorithms were trained they were used to detect attacks form live traffic. Advantage is to produce high accuracy of classification.

2. Support Vector Machine:

Support Vector Machine (SVM) is used to classify the fruit quality. SVM Support vector machines are mainly two class classifiers, linear or non-linear class boundaries.
The idea behind SVM is to form a hyper plane in between the data sets to express which class it belongs to.
The task is to train the machine with known data and then SVM find the optimal hyper plane which gives maximum distance to the nearest training data points of any class.
Steps:
Step 1: Read the test features and trained features. Step 2: Check the all test features of dataset and also get all train features.
Step 3: Consider the kernel.
Step 4: Train the SVM using both features and show the output.
Step 5: Classify an observation using a Trained SVM

Classifier.

E. Dataset

The project was tested by using the KDD CUP 99 DATA SET. The 1998 dataset contains seven weeks of training and also two weeks of testing data. In total, there are 38 attacks in training data as well as in testing data. The refined version of dataset which contains only network data (i.e. Tcpdump data) is termed as KDD dataset.The size of data approx 2 GB.

Results and Discussion

A. Attack Recognition

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.

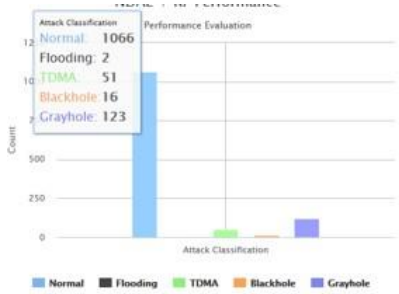

Fig. 2. Performance analysis graph to count the attacks

B. Comparative Results

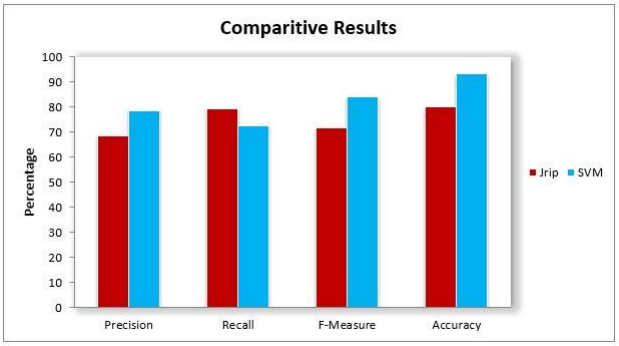

Fig. 3. Comparative Results

Table I Comparative Table

	Jrip	Support vector machine(SVM)
Precision	68.45	78.70
Recall	79.44	72.64
F-Measure	72.11	84.31
Accuracy	80.29	86.26
Execution Time (ms)	435	245

Conclusion

We have discussed the NIDPS's Framework and how it works to detect the intrusion in system.We have discussed the technique used to train machine learning with help of datasets. We have then built Restricted Boltzmann Machine and Deep Belief Network if Training and get Required Classification of intrusions.

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.

[8]. 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.

[9]. 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.

[10]. R. Chitrakar, C. Huang," Selection of candidate support vectors in incremental svm for network intrusion detection", Computers Security 45 (2014) 231241.

[11]. 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.