Filter-based algorithms for implementing an intrusion detecting system during KDD

K.Vivek[1], Mr G.Ananthnath[2]
[1]*Student, Dept.of MCA, KMMIPS, Tirupathi*
[2]*Assistant Professor, Dept. of MCA, KMMIPS,Tirupati*

***Abstract*- Redundant and unsuitable features in data have caused a long-run drawback in network traffic classification. These features not solely slow down the method of classification however also stop a classifier from creating accurate decisions, particularly once coping with big data. In this paper, we tend to propose a mutual data based mostly algorithmic program that analytically selects the optimum feature for classification. This mutual data based feature selection algorithmic program will handle linearly and nonlinearly dependent data options. Its effectiveness is evaluated within the cases of network intrusion detection. an Intrusion Detection System named Least sq. Support Vector Machine based mostly IDS (LSSVM-IDS), is built using the features chosen by our proposed feature choice algorithmic program. The performance of LSSVM-IDS is evaluated using 3 intrusion detection analysis datasets, specifically KDD Cup 99, NSL-KDD and Kyoto 2006+ dataset. The analysis results show that our feature choice algorithm contributes a lot of critical options for LSSVM-IDS to attain higher accuracy and lower process value compared with the state-of the-art methods.**

***Index Terms*- Intrusion detection, Least square support vector machine, Feature selection.**

1. INTRODUCTION

Attacks in the system are becoming frequent and their detection is gaining importance. With the advancement of technology and dependence of human on growing internet and its applications, the safety of information, data flowing over the networks is becoming crucial. In order to prevent this crucial information and to achieve confidentiality, security of networks is one of the critical requirements of the growing network world these days. In the current scenario intrusion detection is mostly human dependent, human analysis is required for detection of intrusion. Systems depend heavily on manual input. Intrusion detection is today important for business organization, system applications and for large number of servers and on-line services running in the system. But, for preventing the data over the network, enhancing the efficiency of the intrusion detection model is also equally important.

There are mainly two types of Intrusion detection system, named as Network based and Host based.

1.Network Based Intrusion Detection System (NIDS):-

The prominence factor of network based intrusion detection system (NIDS) is that at any single instance of time, NIDS can monitor multiple systems in a network in parallel. NIDS finds its best values when each single packet is analyzed that is about to move into the network through the firewall implemented above the network and thus helps in monitoring the information traversing through the network and detects any adversary or intrusion activities.

2.HOST BASED INTRUSION DETECTION SYSTEM (HIDS)

In contrast to NIDS, host based intrusion detection system (HIDS) monitors the activities of an individual host or computer system. The primary focus of host based intrusion detection system is on the operating system activities and events. However, in network systems also, HIDS finds its best values in finding the flow of information and detecting the attacks over the network based on the events occurred within the network.

Computers have been networked together with very large user source and so security has been a vital concern in many areas. With the rapid growth of internet communication and availability of tools to intrude the network, security for network has become indispensable. Current security policies do not sufficiently guard the data stored in the databases. Many other technologies like firewalls, encryption and authorization mechanisms can offer security, but they are still sensitive for attacks from hackers who take advantage of the system flaws.

To protect these systems from being attacked by intruders, a new Intrusion Detection System has been proposed and implemented in this project work, which combines a simple feature selection algorithm and SVM technique to detect attacks. Using KDD cup data set and Data Mining extract the hidden predictive information from large Databases. It is a powerful new technology with great potential that helps companies focus on the most important information in their data warehouses. Data mining can be applied to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Recently, internet has become a part of daily life. The current internets based on information processing systems are prone to different kind of threats which lead to various types of damages resulting in significant losses. Therefore, the importance of information security is evolving quickly. The most basic goal of network security is to develop defensive networking systems which are secure from unauthorized access, using disclosure, disruption, modification, or destruction. Moreover, network security minimizes the risks related to the main security goals like confidentiality, integrity and availability.

3. INTRUSION DETECTION FRAMEWORK BASED ON LEAST SQUARE SUPPORT VECTOR MACHINE

The framework of the proposed intrusion detection system is depicted in Figure 1. The detection framework is comprised of four main phases: (1) data collection, where sequences of network packets are collected, (2) data preprocessing, 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 LSSVM, and (4) attack recognition, where the trained classifier is used to detect intrusions on the test data.

Support Vector Machine (SVM) is a supervised learning method. It studies a given labeled dataset and constructs an optimal hyper plane in the corresponding data space to separate the data into different classes. Instead of solving the classification problem by quadratic programming, Suykens and Vandewalle suggested re-framing the task of classification into a linear programming problem. They named this new formulation the Least Squares SVM (LS-SVM). LS-SVM is a generalized scheme for classification and also incurs low computation complexity in comparison with the ordinary SVM schem. One can find more details about calculating LS-SVM in Appendix B. The following subsections explain each phase in detail.

1.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 IDS. To provide the best suited protection for the targeted host or networks, this study proposes a network based IDS to test our proposed approaches. The proposed IDS run on the nearest router to the victim(s) and monitor the inbound network traffic. During the training stage, the collected data samples are categorized with respect to the transport/Internet layer protocols and are labeled against the domain knowledge. However, the data collected in the test stage are categorized according to the protocol types only.

2.Data Preprocessing:-

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.

A)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.

B)Data normalization:-An essential step of data preprocessing 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. 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.

C)Feature selection:-Even though every connection in a dataset is represented by various features, not all of these features are needed to build 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 proposed 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 are depicted in Table 1 [a-c], 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 selection algorithm is applied for the aforementioned classes. The selected features are shown in Table 3.

3.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.

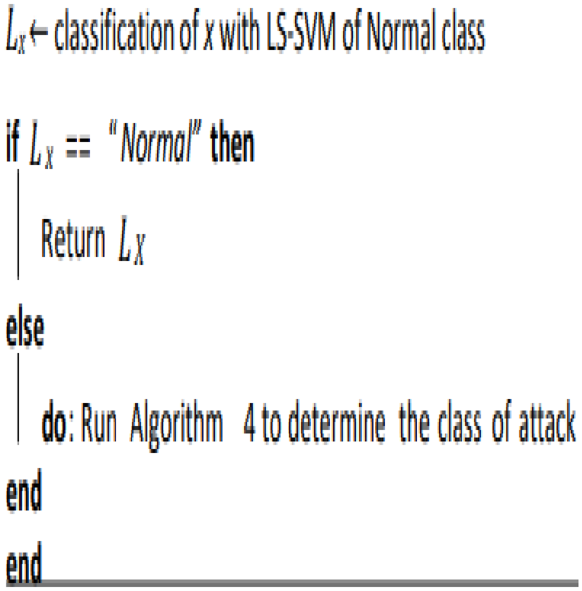
4.Attack Recognition:-

In general, it is simpler to build a classifier to distinguish between two classes than considering multiclass 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. However, to deal with a problem having more than two classes, there are two popular techniques: “One-VsOne” (OVO) and “One-Vs-All” (OVA). Given a classification problem with M classes (M > 2), the OVO approach on the one hand divides an M-class problem into $\frac{M*(M-1)}{2}$ binary problems. Each problem is handled by a separate binary Algorithm 3 Intrusion detection based on LS-SVM {Distinguishing intrusive network traffic from normal network traffic in the case of multiclass}

Input: LS-SVM Normal Classifier, selected features (normal class), an observed data item x

Output: Lx - the classification label of x

Begin:



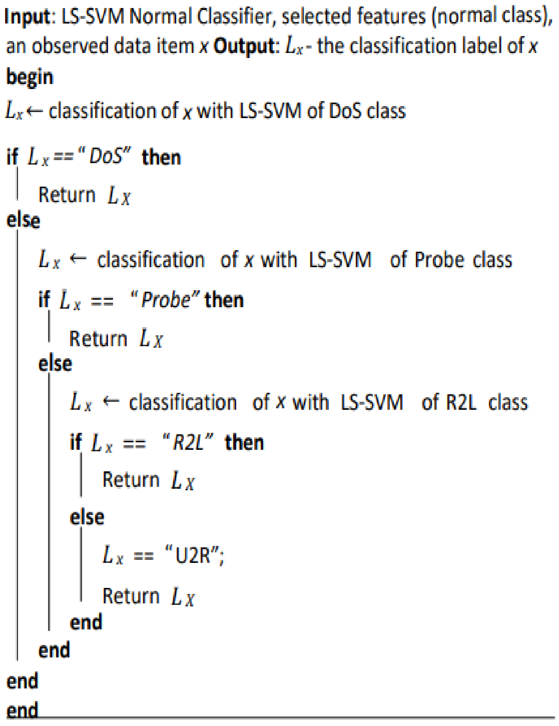
```
$L_x$ ← classification of $x$ with LS-SVM of Normal class
if $L_x$ == "Normal" then
    Return $L_X$
else
    do: Run Algorithm 4 to determine the class of attack
end
end
```

Classifier, which is responsible for separating the data of a pair of classes.

The OVA approach, on the other hand, divides an Mclass problem into M binary problems. Each problem is handled by a binary classifier, which is responsible for separating the data of a single class from all other classes. Obviously, the OVO approach requires more binary classifiers than OVA. Therefore, it is more computationally intensive. Rifkin and Klautau demonstrated that the OVA technique was preferred over OVO. As such, the OVA technique is applied to the proposed IDS to distinguish between normal and abnormal data using the LS-SVM method.

After completing all the aforementioned steps and the classifier is trained using the optimal subset of features which includes the most correlated and important features, the normal and intrusion traffics can be identified by using the saved trained classifier. The test data is then directed to the saved trained model to detect intrusions. Records matching to the normal class are considered as normal data, and the other records are reported as attacks. If the classifier model confirms that the record is abnormal, the subclass of the abnormal record (type of attacks) can be used to determine the record's type. Algorithms 1 and Algorithm 2 describe the detection processes.

Algorithm 2 Attack classification based on LS-SVM



```
Input: LS-SVM Normal Classifier, selected features (normal class),
an observed data item $x$ Output: $L_x$- the classification label of $x$
begin
$L_x$ ← classification of $x$ with LS-SVM of DoS class
if $L_x$ == "DoS" then
    Return $L_X$
else
    $L_x$ ← classification of $x$ with LS-SVM of Probe class
    if $L_x$ == "Probe" then
        Return $L_X$
    else
        $L_x$ ← classification of $x$ with LS-SVM of R2L class
        if $L_x$ == "R2L" then
            Return $L_X$
        else
            $L_x$ == "U2R";
            Return $L_X$
        end
    end
end
end
```

4. CONCLUSION

In this paper, a supervised filter-based feature choice algorithm has been proposed, specifically versatile Mutual info Feature selection (FMIFS).FMIFS is then combined with the LSSVM methodology to build an IDS. LSSVM may be a least square version of SVM that works with equality constraints rather than inequality constraints within the formulation designed to solve a set of linear equations for classification issues instead of a quadratic programming problem. The planned LSSVMIDS + FMIFS has been evaluated using 3 accepted intrusion detection datasets: KDD Cup 99, NSL-KDD and Kyoto 2006+ datasets. The performance of LSSVMIDS + FMIFS on KDD Cup test data, KDDTest+classification performance in terms of classification accuracy, detection rate, false positive rate and Fmeasure than some of the existing detection approaches.

REFERENCES

[1] S. Pontarelli, G. Bianchi, S. Teofili, Traffic-aware design of a highspeedfpga network intrusion detection system, Computers, IEEETransactions on 62 (11) (2013) 2322–2334.

[2] B. Pfahringer, Winning the kdd99 classification cup: Bagged boosting, SIGKDD Explorations 1 (2) (2000) 65–66.

[3] I. Levin, Kdd-99 classifier learning contest: Llsoft'sresultsoverview, SIGKDD explorations 1 (2) (2000) 67–75.

[4] D. S. Kim, J. S. Park, Network-based intrusion detection withsupport vector machines, in: Information Networking, Vol. 2662, Springer, 2003, pp. 747–756.

[5] A. Chandrasekhar, K. Raghuveer, An effective technique for intrusiondetection using neuro-fuzzy and radial svmclassifier,in: Computer Networks & Communications (NetCom), Vol. 131,Springer, 2013, pp. 499–507.

[6] S. Mukkamala, A. H. Sung, A. Abraham, Intrusion detection usingan ensemble of intelligent paradigms, Journal of network andcomputer applications 28 (2) (2005) 167–182.

[7] A. N. Toosi, M. Kahani, A new approach to intrusion detectionbased on an evolutionary soft computing model using neurofuzzyclassifiers, Computer communications 30 (10) (2007) 2201–2212.

[8] Z. Tan, A. Jamdagni, X. He, P. Nanda, L. R. Ping Ren, J. Hu,Detection of denial-of-service attacks based on computer visiontechniques, IEEE Transactions on Computers 64 (9) (2015) 2519–2533.

[9] A. M. Ambusaidi, X. He, P. Nanda, Unsupervised feature selectionmethod for intrusion detection system, in: InternationalConference on Trust, Security and Privacy in Computing andCommunications, IEEE, 2015.

[10] A. M. Ambusaidi, X. He, Z. Tan, P. Nanda, L. F. Lu, T. U. Nagar, A novel feature selection approach for intrusion detection dataclassification, in: International Conference on Trust, Security and Privacy in Computing and Communications, IEEE, 2014, pp. 82– 89.

[11] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks5 (4) (1994) 537– 550.

[12] F. Amiri, M. RezaeiYousefi, C. Lucas, A. Shakery, N. Yazdani,Mutual information-based feature selection for intrusion detection systems, Journal of Network and Computer Applications 34 (4)(2011) 1184–1199.

[13] A. Abraham, R. Jain, J. Thomas, S. Y. Han, D-scids: Distributedsoft computing intrusion detection system, Journal of Networkand Computer Applications 30 (1) (2007) 81–98.

[14] S. Mukkamala, A. H. Sung, Significant feature selection using computational intelligent techniques for intrusion detection, in: Advanced Methods for Knowledge Discovery from Complex Data, Springer, 2005, pp. 285–306.

[15] S. Chebrolu, A. Abraham, J. P. Thomas, Feature deduction and ensemble design of intrusion detection systems, Computers &Security 24 (4) (2005) 295–307.