Improving Accuracy for Intrusion Detection using Support Vector Machine with Feature Reduction

[1]Gangaprasad G. Ghungre, [2]Prof. T. A. Rane
[1]Student of Pune Institute of Computer technology, [2]Asst. Prof. of of Pune Institute of Computer technology
[1]Information Technology,
[1]Pune Institute of Computer technology, Pune,India
Email – [1]ghungre.g@gmail.com, [2]tarane@pict.edu

Abstract: *21st century has witnessed internet growth by heaps and bounds. Technology has marked its existence in each and every field ranging from science, commerce to defence, entertainment, sports etc. With the massive increase in the use of internet, all the sensitive and confidential data is available on the network. This gives the scope to the intruders to have unprivileged access to this data and exploit it in such a way that it causes financial and reputational to that respective organization or person. Cyber–attacks can be broadly classified into four categories i.e. DOS, Probe, U2R, R2L. As the popularity of internet is increasing, the risk of network attack growing. Intrusion detection helps us to detect an avoid these cyber attacks. Cyber security analysts still desire much more accuracy by monitoring security events. From more than decade, different methods of machine learning are being applied for intrusion detection(ID). Aim is to improve accuracy for intrusion detection(ID) using Support Vector Machine(SVM) with Feature Reduction technique. For this purpose, we are using NSL-KDD dataset.*

Key Words: Intrusion Detection (ID), NSL-KDD Data set, Support Vector Machine(SVM)

1. INTRODUCTION:

A. Intrusion Detection Syatem(IDS)

Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of intrusion. Intrusions are defined as attempts to compromise the confidentiality, integrity and availability of network. This security mechanism can be implemented using an Intrusion Detection System (IDS) which can be describe as a collection of software or hardware device able to collect, analyze and detect any unwanted, suspicious or malicious traffic either on a particular computer host or network (1).

An attack generally falls into one of four categories:

- Denial-of-Service(DoS): Attackers tries to make a network recourse unavailable or too busy. e.g. are smurf, neptune, back, teardrop, pod and land.
- Probe: Attackers tries to gain information about the target host. Port Scans or sweeping of a given IP-address range typically fall in this category (e.g. saint, ipsweep, portsweep and nmap).
- User-to-Root(U2R): Attackers has local access to the victim machine and tries to gain super user privileges. e. g. are buffer overflow, rootkit, perl.
- Remote-to-Local(R2L): Attackers does not have an account on the victim machine, hence tries to gain access. e. g. are guess_passwd, ftp_write, multihop,phf, spy, imap.

B. Support Vector Machines (SVM)

Support Vector Machines (SVMs), also known as Support Vector Networks are supervised learning method used in machine learning, with associated learning algorithms that analyze data and identify pattern. It is also used for categorization and regression analysis. Support vector machines (SVM) classify data with different class labels by determining a set of support vectors that are members of the set of training inputs that outline a hyper-plane in the feature space. SVM provides a generic mechanism which fits the hyper plane surface to the training data using a kernel function (2).

C. Feature Selection

Feature selection is important to improve the efficiency of data mining algorithms. Most of the data include irrelevant, redundant or noisy features. Feature selection is process of selecting a subset of original features according to certain criteria. It is an important and frequently used technique in data mining for dimension reduction. It reduces the number of features, removes irrelevant, redundant or noisy features, and brings about palpable effects for applications: speeding up a data mining algorithm, improving learning accuracy, and leading to better model comprehensibility (3).

There are mainly three common approaches for feature reduction: Filter method, Wrapper method and Embedded method. Wrapper method uses the intended learning algorithm itself to evaluate the usefulness of features, while filter evaluates features according to heuristics based on general characteristics of the data. The wrapper approach is generally considered to produce better feature subsets but runs much more slowly than a filter (4). Feature selection done using Principle Component Analysis and use of Information Gain(IG).

a. Principal Component Analysis(PCA)

Principal Component Analysis(PCA) is a useful statistical technique that has found application in fields such as face recognition and image compression, and is a common technique for finding patterns in data of high dimension. The entire subject of statistics is based on around the idea that you have this big set of data, and you want to analyze that set terms of the relationships between the individual points in that set. The goal of PCA is to reduce the dimensionality of the data while retaining as much as possible of the variation present in the original dataset. It is a way of identifying patterns in data, and expressing the data in such a way as to high light their similarities and differences (5)(6).

b. Information Gain (IG)

The IG evaluates attributes by measuring their information gain with respect to the class. It discretizes numeric attributes first using MDL based discretization method (7).

2. LITERATURE REVIEW:

Today the world is experiencing the thrill and miracles of technology. Growth in information technology has affected every aspect of life including communication, transport, society and almost everything. The basic system for detecting any type of attacks and threats in cyber space at the network level by examining network traffic is termed as Network Intrusion Detection System(NIDS).

Here we are considering various papers related to topic. Here we are covering previews work for DOS, IDS and feature reduction.

A. Intrusion Detection System(IDS)

In this paper, for building an Intrusion Detection System (IDS) use of Filter based selection process, in this work intends to explore a mean of selecting optimal feature from feature space regardless of the type of correlation between them. Feature selection method based on mutual information and the framework for IDS based on Least Square SVM, in this firstly collect the data and then it processes after use of classifier training, where the model for classification is trained using LS-SVM and lastly attack recognition. In the experimental result shows that detection model combined with Flexible Mutual Information Feature Selection (FMIFS) has achieved an accuracy rate of 99.79%, 99.91% and 99.71% for KDD Cup99, NSL-KDD and Kyoto 2006+ dataset respectively (5). Mukkamala et al. (22) investigated the possibility of assembling various learning methods, including Artificial Neural Networks (ANN), SVMs and Multivariate Adaptive Regression Splines (MARS) to detect intrusions. They trained five different classifiers to distinguish the normal traffic from the four different types of attacks. They compared the performance of each of the learning methods with their model and found that the ensemble of ANNs, SVMs and MARS achieved the best performance in terms of classification accuracies for all the five classes. Muhammad Imran et al. (23) applied Linear Discriminant Analyis (LDA) and Genetic Algorithm for feature selection and further implemented Radial Basis Function for feature classifier. He applied cross validation on 20% of NSL KDD training dataset for training and testing. Desta Haileselassie Hagos et al. (24) In this paper, address the problem of an actual feature selection for IDS to find attack categories in a network through cross-validated regularized ML techniques and an artificial neural network feature ranking method. Selecting the most relevant actual features improves the detection quality for many algorithms that are based on learning techniques. They focus on the analyze security attacks by exploring the contribution of the 41 widely used actual input features and selecting the most contributory ones in effectively identifying anomalies in a network with respect to the attack categories. For feature selection they ranked the actual input features into strongly contributing, low contributory and irrelevant using a combination of feature selection filters and wrapper methods by carrying out comparisons with previous works. We investigate the most important features in identifying well-know security attacks by using SVMs and regularized method with LASSO.

The main contributions of our paper are Performed extensive simulation results are compared feature ranking using both *two-stage*1 approach using SVM and *one-stage*2 approach using LASSO. They found that LASSO provides comparable results at a lower computational cost.

B. Denial of service(DOS)

In this paper, use of penetration testing technique for simulating DoS attacks in order to verify the level of security systems implemented in the government network of Kosovo. Simulate different scenarios for different types of Firewalls and they came up with important suggestions for improving protective security system in the Government network, and assume that the attacker is attacking from local network and from Internet (outside the network). Showing that how DoS attacks affect up-and-running web services, which are located in a test server inside this network. For this purpose, use of penetration testing techniques. Penetration testing is a method that simulates an attack in order to verify the security holes of a particular system. This test can be performed using hardware or software tools, but also through social engineering. The main purpose of this method is to examine the behavior of a particular system during an attack from inside or outside an Organization. They use different testing module like white box, Gray box, Black box , Depending on the information that pen-tester has regarding the domain of testing, penetration testing can be done from inside or outside network of the organization(19).

Zhiyuan Tan et al. (20) in this paper Denial of Service attack detection based on principle of Multivariate Correlation Analysis (MCA) techniques and anomaly based detection system with capabilities of accurate characterization for traffic behaviors and detection of known and unknown attacks respectively. A statistical normalization technique is used to eliminate the bias from the raw data. The proposed DoS detection system is evaluated using KDD Cup 99 dataset (21) and outperforms the state-of the- art systems. In the MCA framework, in which Triangle Area Map Generation module is applied to extract the correlations between two distinct features within each record. For Detection Mechanism a threshold based anomaly detector used and whose normal profiles are generated using purely legitimate network traffic records. The evaluation of the proposed DoS attack detection system is conducted on KDD Cup 99 dataset. For the performance they use 10-fold cross validation along with change of threshold, the rate of correct classification of the Normal records rise from 98.74% to 99.47%.

C. Feature selection

In this work investigates the effectiveness and the feasibility feature reduction technique on Back Propagation Neural Network classifier. They had performed various experiments on KDD CUP 1999 dataset and recorded Accuracy, Precision, Recall values. In this work, they had done Basic, N-Fold Validation and Testing comparisons on reduced dataset with full feature dataset. Basic comparison clearly shows that the reduced dataset outer performs on size, time and complexity parameters. Experiments of N-Fold validation show that classifier that uses reduced dataset, have better generalization capacity. During the testing comparison, found both the datasets are equally compatible. All the three comparisons clearly show that reduced dataset is better or is equally compatible, and does not have any drawback as compared to full dataset. In this experiment shows that usage of such reduced dataset in BPNN can lead to better model in terms of dataset size, complexity, processing time and generalization ability (9).

D. Feature Reduction

Most of the proposed research system could effectively utilize feature selection process to improve detection rate of their system and minimize considerably the false alarm rate. Research usually missed to detect new intrusions, especially when the intrusion mechanism used differed from the previous intrusion.

In 2009, Shi-Jinn (8) works revealed that not all research carried out feature selection before they trained their classifier, however based on (9) (10), these processes take a significant part to different types of intrusion identification and features can be excluded without the performance of the IDS to be dropped. Juan Wang et al., in their work (11) proposed a decision tree based algorithm for intrusion detection, even if during their experiments the algorithm was achieving a good detection accuracy, the error rate was remaining identical.

Back in 2010, Farid et al. (12), used a decision tree based learning algorithm to retrieve important features set from the training dataset for intrusion detection. Their techniques found relevant features using a combination of ID3 and C4.5 decision tree algorithms. They assigned a weight value to each feature. The weight is determined where the minimum depth of the decision tree at which each feature is checked inside the tree and the weights of features that do not appear in the decision tree are allocated a value of zero. Ektefa et al. (13), used different data mining method for intrusion detection and they found that the decision tree classifier was performing better than the SVM learning algorithm.

Geetha Ramani et al. (14) used in their paper in 2011, a statistical method for analyzing the KDD 99 dataset. They identified the important features by studying the internal dependences between features.

In their paper proposed in 2012, S. Mukherjee and N. Sharma (15) designed a technique called Feature- Vitality Based Reduction Method (FVBRM) using a Naïve Bayes classifier. FVBRM identifies important features by using a sequential search approach, starting with all features, one feature is removed at a time until the accuracy of the classifier reaches some threshold. Their method shows an improvement of the classification accuracy but takes more time and still complex when detecting the U2R attacks.

In 2013, support vector machine classifier was used by Yogita B. Bhavsar et al. (16), for intrusion detection using the NSL KDD dataset. The drawback with this technique is the extensive training time required by the classifier, so to reduce the time, they applied a radial basis function (RBF) to reduce the extensive time.

In 2014, O. Y. Al-Jarrah et al. (17), used an ensemble of decision-tree based voting algorithm with forward selection / backward elimination feature raking techniques using a Random Forest Classifier. Their method shows an improvement of detection accuracy when selected important features and it can be suitable for large-scale network.

N. G. Relan and D. R. Patil (18) in their papers have tested two decision tree approach to classify attacks using the NSL KDD dataset. They have found that the C4.5 with pruning offers better accuracy than the C4.5 without pruning and it was necessary to reduce the number of features because using all features degrades the performance of the classifier also its time consuming.

3. METHOD:

Existing System

For the identification and classification of network attacks use sevreal techniques like Random Forest, Decision Tree, Naïve Bayes, Neural Network ,Support Vector Machine etc.are used but not getting more accurate accuracy.

Proposed System

In this paper a Support Vector Machine algorithm with Feature reduction process that is Principal of Component analysis(PCA) is proposed for the Intrusion Detection System(IDS)to detect anomaly from NSL-KDD dataset. Feature reduction process can be viewed as a pre-processing step which removes distracting variance from a dataset, so that classifiers can perform better. In this proposed algorithm PCA transform used for dimensionality reduction which is commonly used step, especially when dealing with high dimensional space of features. PCA-based approaches improve system performances and to identify any unknown attacks (19). The performance of the different kernels based approach on the basis of their accuracy in term of false positive rate and precision.

A. Mathematical Flow:

Let 'S' be the system comparing object and function

S={S, E, I, O ,success, failure}

S=Start of system

E=Result

S={s1, s2, s3}

Where,

s1=Data set

s2=Feature reduction

s3=Classification

Input I ={data set}

Output O ={Attacks detected and classified}

Success ={Normal, Abnormal}

Failure ={Not detected any Attacks}

Proposed Methodologies

The contrast of performance of numrous intrusion detection system is done by the nominated feature which is done by the feature selection algorithm and the classification of attack is done using machine learning algorithms.For feature reduction the Information Gain(IG) based feature selection algorithm is used in our framework.

A. Dataset

To resolve some issues found in the previous KDD 99, an improved version was created, the NSL KDD dataset. The reason behind the use of this dataset has been reported at among them the following are relevant to mention: Elimination of redundant records in the training set will help our classifier to be unbiased towards more frequent records.

B. Data Pre-processing

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

b.1Data 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 NSL-KDD 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.2 Data normalization.

An essential step of data pre-processing 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 favour of features with greater values is eliminated from the dataset.

C. Features Selection

Feature selection is used to eliminate the redundant and irrelevant data. It is a technique of selecting a subset of relevant features that fully represents the given problem alongside a minimum deterioration of presentation (21), two possible reasons were analyzed why it would be recommended to restrict the number of features:

Firstly, it is possible that irrelevant features could suggest correlations between features and target classes that arise just by chance and do not correctly model the problem. This aspect is also related to over-fitting, usually in a decision tree classifier. Secondly, a large number of features could greatly increase the computation time without a corresponding classifier improvement. There are three main categories of feature selection techniques Filter Method, Wrapper Method and Embedded Method.

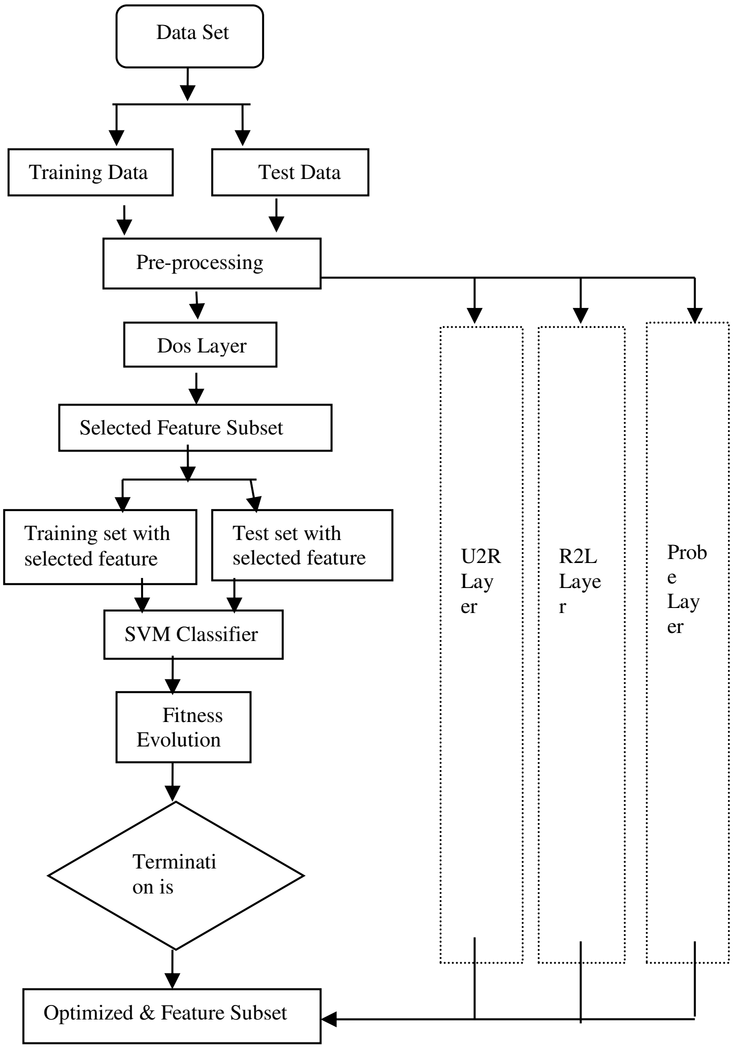

Fig.1 Architecture diagram for IDS

4. CONCLUSION:

In this proposed work, feature selection methods using Principle Component Analysis (PCA) and Information Gain(IG). As far as we know, previous researches did not perform simultaneous feature selection and parameters optimization for support vector machines. We conducted experiments to calculate the classification accuracy of the proposed approach.

In this research work, evaluated the performance of different kernels for SVM used in IDS. The performances of the different kernels based approach has been observed on the basis of their accuracy in terms of false positive rate and precision. The results indicate that the performance of the SVM classification depends mainly on the types of kernels and their parameters. The obtained results justify the motivation of this work that only a single kernel cannot be considered for SVM used in IDS to achieve the optimal performance. Research in intrusion detection using SVM approach is still demanding due to its better performance.

REFERENCES:

1. Arianit Maraj, Genc Jakupi, Ermir Rogova and Xheladin Grajqevci, (2017):Testing of Network Security Systems Through DoS Attacks, BAR, MONTENEGRO.
2. LI Hui, GUAN Xiao-hong, ZAN Xin, (2003):Network intrusion detection based on support vector machine, Journal of computer research and development,pp. 800-807.
3. Liu, H., Motoda, H., Setiono, R., & Zhao, Z., (2010):Feature Selection: An Ever-Evolving Frontier in Data Mining,*Journal of Machine Learning Research-Proceedings Track 10*, pp.4-13.
4. Kim, Yong, W. Nick Street, and Filippo Menczer., (2003):Feature selection in data mining, *Data mining: opportunities and challenges Vol.3, No.9*, 2003, pp.80-105.
5. Aditi Nema, Dr. Basant Tiwari, Dr. Vivek Tiwari, (2016): Improving Accuracy for Intrusion Detection through Layered Approach Using Support Vector Machine with Feature Reduction, ACM.ISBN978-1-4503-4278-0/16/03, DOI:10.1145/2909067.2909100
6. Lindsay I Smith A tutorial on Principal Components Analysis February 26,2002.
7. j.Han ,M Kamber, (2001): Data mining : Concepts and Techniques, Morgan Kauffmann Publishers.
8. C. F. Tsai, et al., (2009): Intrusion detection by machine learning: A review, vol. 36, pp. 11994-12000.
9. V. Bolón-Canedo, et al., (2011):Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset, vol. 38, pp. 5947-5957, 2011.
10. F. Amiri, et al., (2011):Improved feature selection for intrusion detection system.
11. Juan Wang, Qiren Yang, Dasen Ren, (2009): An intrusion detection algorithm based on decision tree technology.
12. Dewan Md. Farid, Nouria Harbi, and Mohammad Zahidur Rahman, (2010):Combining Nave Bayes and Decision Tree for Adaptive Intrusion Detection Vol. 2, No. 2, pp. 12-25.
13. Ektefa M, Memar S, Sidi F, Affendey L., (2010):Intrusion detection using data mining techniques, doi:10.1109/infrkm.2010.5466919.
14. Geetha Ramani R, S.SivaSathya, SivaselviK, (2011): Discriminant Analysisbased Feature Selection in KDD Intrusion Dataset, VoI.31,No.ll.
15. S. Mukherjee and N. Sharma, (2012):Intrusion Detection using Naive Bayes Classifier with Feature Reduction, vol. 4, pp. 119–128.
16. Bhavsar Y. B, Waghmare K. C. (2013):Intrusion Detection System Using Data Mining Technique: Support Vector Machine, Vol.3, Issue 3, pp.581-586(2013).
17. O. Y. Al-Jarrah, a. Siddiqui, M. Elsalamouny, P. D. Yoo, S. Muhaidat, and K. Kim, (2014): Machine-Learning-Based Feature Selection Techniques for Large-Scale Network Intrusion Detection, pp. 177–181.
18. N. G. Relan and D. R. Patil, (2015):Implementation of Network Intrusion Detection System using Variant of Decision Tree Algorithm , pp. 3–7.
19. Prof. Bhavin Shah, Prof. Bhushan H Trivedi, (2017):Reducing Features of KDD CUP 1999 Dataset for Anomaly Detection Using Back Propagation Neural Network, DOI 10.1109/ACCT.2015.131
20. Zhiyuan Tan, Aruna Jamdagni, Xiangjian He, Priyadarsi Nanda, and Ren Ping Liu, (2009): A System for Denial-of-Service Attack Detection Based on Multivariate Correlation Analysis,DOI 10.1109/TPDS.2013.146.
21. S. J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan, (2000): Costbased modeling for fraud and intrusion detection: results from the JAM project, The DARPA Information Survivability Conference and Exposition , Vol.2, pp. 130-144, 2000.
22. Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi Nanda, and Zhiyuan Tan, (2014): Building an intrusion detection system using a filter-based feature selection algorithm IEEE Transactions on computer, DOI 10.1109/TC.2016.2519914.
23. S. Mukkamala, A. H. Sung, A. Abraham, (2005): Intrusion detection using an ensemble of intelligent paradigms, Journal of network and computer applications 28 (2) 167–182.
24. Desta Haileselassie Hagos, Anis Yazidi, Oivind Kure, Paal E.Engelstad, (2017):Enhancing Security Attacks Analysis using Regularized Machine Learning Techniques, DOI 10.1109/AINA.2017.19.