IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Multi-Attacks Detection in Distributed System using Machine Learning

P. Patil[1*], T. Bagwan[2], S. Kulkarni[3], C. Lobo[4], S.R. Khonde[5]

[1,2,3]Department of Computer Engineering, Modern Education Society's College of Engineering, SPPU, Pune, India

[*]*Corresponding Author: prajaktapatil314@gmail.com, Tel.: +91-8975627276*

Available online at: www.ijcseonline.org

Accepted: 14/Jan/2019, Published: 31/Jan/2019

Abstract— Intrusion compromises a computer by breaking its security and thereby the computer enters into an insecure state. If such an event takes place, the computer becomes vulnerable to several attacks. These attacks aim to obtain information about the target computer and the information so obtained can be used to conduct fraudulent activities. It is difficult to prevent an intrusion into the system. However, if these computer intrusions are detected in time, the administrator can be informed and necessary actions can be taken at early stages. Previous Intrusion detection system (IDS) utilized several features to detect various malicious activities. However, these IDS methods only detect specific attack. They fail when the attacks are combined. For this purpose, we propose an Intrusion Detection System in distributed environment to mitigate the individual and combination routing attacks. This paper explains the method we used to generate such a system. Our proposed system of Intrusion Detection uses feature selection techniques to determine significant features, along with the best classification method will distinguish between an attack and non-attack. We aim to increase detection accuracy and reduce false alarm rate. NSL-KDD dataset has been used to train our model. The paper also explains related work done in this field and briefly explains the network attacks and the dataset.

Keywords— IDS, Intrusion Detection System, Multiple attacks, Machine Learning, Network Security.

I. INTRODUCTION

There has been a rapid increase in Internet usage in the recent past that has led to an increase in malicious network activity. With the advent of new technologies each day and widespread of computers (from personal computers to embedded systems), security has become a very important issue. To name a few Attacks like Probe, DoS (Denial of Service), U2R (User-to-Root), R2L (Remote-to-Local) have become a great deal of concern to every computer in the network. Such attacks compromise the security of the computer and obtain access to sensitive data. Hence, security of any network is a high priority issue which must be taken care of. Various Intrusion Detection Systems (IDS) exist which help identify threats in the system. An Intrusion detection system is a machine or software that monitors the traffic in a network and on detection of a malicious packet, informs the user or a specific acting unit which can take further action and avoid the malicious packet from entering the network. Earlier IDS developed were capable of detecting only individual attacks but failed when the attacks were combined.

Also, with Machine Learning becoming pervasive in the computer world, it sets its foot into the area of Network Security as well. Hence, we could make full use of it and create a system that could provide a secure environment for the users in a network.

Our aim is to create such a system. The above mentioned issues motivated us to select this project. Our Aim is to create an IDS which incorporates the methodologies of Machine Learning to identify the individual and combination routing attacks in the network correctly with very less number of misclassifications which otherwise would go unidentified in traditional Intrusion Detection Systems.

The paper is organized as follows: Section 2 is about the project scope. Section 3 provides literature review. In Section 4 and Section 5 dataset, attacks and data pre-processing is discussed respectively. Feature Selection techniques are explained briefly in Section 6. In Section 7, classifiers are discussed. Finally, the conclusions are drawn in Section 8. And then acknowledgements in Section 9 and references in Section 10.

II. PROJECT SCOPE

The aim is propose an IDS in distributed environment to mitigate the individual and combination routing attacks. We will be using feature selection techniques to determine significant features for intrusion detection by identifying the best classification method that will be able to distinguish between an attack and non-attack so as to increase detection accuracy and reduce false alarm rate. On Detection of an attack the user or network administrator will be alarmed

either through a computer software appropriate actions can be taken.

Before starting the implementation an appropriate dataset must be selected to train our classifiers. We have selected the NSL – KDD as our dataset. The next step is to select the feature selection techniques and then select the classifiers. Random forest, Naïve Bayes and J48 are the three classifiers that we have selected. After selection of classifiers, modelling and training is performed. The best classification method is selected so as to increase detection accuracy and reduce false alarm rate.

III. LITERATURE REVIEW

Mohamad Nazrin Napiah, Mohd Yamani Idna Idris, Roziana Ramli, Ismail Ahmedy developed an IDS for three attacks and their combination. They utilized Best First and Greedy Step-Wise with Correlation based Feature Selection to determine significant features needed for the intrusion detection. These features were then tested using six machine learning algorithms to find the best classification method that was able to distinguish between an attack and non-attack.
Among six algorithms (MLP, SVM, J48, Naive Bayes, Logistic, and Random Forest), J48 algorithm was chosen for the proposed CHA-IDS as J48 showed the best performance among all algorithms in detecting the routing attacks. The three types of combination attacks considered were hello flood, sinkhole, and wormhole. The results considered were in term of accuracy of detection, energy overhead and memory consumption with the prior 6LoWPAN-IDS implementation such as SVELTE and Pongle IDS. The results showed that CHA-IDS performs better with 99 percent true positive rate and consumed low energy overhead and memory. The consumption of energy and memory of CHA-IDS were 5840mW and 44.9kB respectively. The IDS effectively detected both individual and new anomaly attack that created by combination of the routing attacks [1].

Preeti Aggarwala, Sudhir Kumar Sharma presented the analysis of KDD data set with respect to four classes which are Basic, Content, Traffic and Host in which all data attributes can be categorized. The analysis was done with respect to two prominent evaluation metrics, Detection Rate (DR) and False Alarm Rate (FAR) for an IDS. Random Tree algorithm, a tree based classifier was selected for simulation. WEKA Tool was used for analysis. The results were analyzed to study dominance of each class of attributes in improving the Detection Rate and minimizing the False Alarm Rate [2].

Hee-su Chae, Byung-oh Jo, Sang-Hyun Choi, Twae-kyung Park main purpose was to identify important selected input features in building IDS that is computationally efficient and effective. The classifier decision tree algorithm was used for evaluating feature reduction method. They proposed a method AR (Attributed Ratio) which is a new method for giving importance of classes and compared with GR, IG and CFS. They proved that by using only 22 features out of 41, an accuracy of 99.79 can be achieved [3].

Shailesh Singh Panwar, Dr. Y. P. Raiwani applied data reduction algorithms on NSL-KDD dataset. The output of each data reduction algorithm was given as an input to two classification algorithms i.e. J48 and Naive Bayes. The main was to find out which data reduction technique proves to be useful in enhancing the performance of the classification algorithm. Results were compared on the bases of accuracy, specificity and sensitivity [4].

Harvinder Pal Singh Sasan and Meenakshi Sharma built a hybrid misuse intrusion detection model to find attacks on system to improve the intrusion detection. The model contains the advantage of feature selection and machine learning techniques with misuse detection. The proposed hybrid model for intrusion detection system assumes that higher number of features in the network needs not to be considered to achieve high accuracy. So, the proposed model was implemented over 29 features with the success rate of 88.23% [5].

L.Dhanabal, Dr. S.P. Shantharajah analysed the NSL-KDD data set and used to study the effectiveness of the various classification algorithms in detecting the anomalies in the network traffic patterns. They have also analysed the relationship of the protocols available in the commonly used network protocol stack with the attacks used by intruders to generate anomalous network traffic. The analysis was done using classification algorithms available in the data mining tool WEKA. The authors concluded that most of the attacks are launched using the inherent drawbacks of the TCP protocol [6].

Arjunwadkar Narayan M. and Thaksen J. Parvat proposed an Intrusion Detection System with Machine Learning model Combining Hybrid Classifiers i.e. Naive Byes classifier and C4.5 classifier. They have presented off-line intrusion detection system model. A preprocessor was used to reduce the dimension of feature vectors and shorten training time. The proposed hybrid intrusion detection method was evaluated by conducting experiments with the NSL-KDD data set, which is a modified version of well-known KDD Cup 99 data set [7].

IV. DATASET AND ATTACKS

A. NSL-KDD

The NSL-KDD dataset has total 42 attributes. This data set is an improvement over KDD99 data set from which duplicate instances were removed to get rid of biased classification results. This data set has number of versions available, out of

which 20% of the training data is used which is identified KDDTrain+20Percent with a total number of 25192 instances. The test data set is identified by the name KDDTest+ and has a total of 22544 instances. Different configurations of this data set are available with variation in number of instances but the number of attributes in each case is 42. The attribute labeled 42 in the data set is the class attribute which indicates whether a given instance is a normal connection instance or an attack. One of the advantages of NSL-KDD dataset is that the quantity is reasonable enough to be trained as a whole by most algorithm. Also, both the 22 attacks in the training set and 39 attacks in the test set can be further categorized into four broader groups which are Dos, Probe, U2R and R2L.

B. Types of Attacks

1) Denial of Service (DoS)

Denial-of-service (DOS) attacks have the goal of limiting or denying service(s) provided to a user, computer or network. An attacker tries to prevent legitimate users from using a service by making it unavailable by overloading the server with too many requests to be handled. For example, SYN flood, Smurf and teardrop.

2) Probe

Probing or surveillance attacks have the goal of gaining knowledge of the existence or configuration of a computer system or network. Port scans or sweeping of a given IP address range is typically used in this category like IPsweep. Attacker scans the network with the aim of exploiting a known vulnerability.

3) Remote-to-Local (R2L)

In this, the attacker tries to gain local access to unauthorized information through sending packets to the victim machine. For example, password guessing attack.

4) User-to-Root (U2R)

In this, the attacker gains root access to the system using normal user account to exploit vulnerabilities. For example, buff er overflow attacks.

V. DATA PREPROCESSING

Data preprocessing is the conversion of data into usable and desired form. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data Preprocessing is a proven method to resolve such issues. . The data from NSL-KDD dataset is preprocessed before it is used for further use. Preprocessing such as checking and removal of missing values, converting nominal data into numeric form and normalizing values into particular range is done. Data normalization is extremely important for training classifier using NSL-KDD dataset as range of value for each feature varies a lot. If data is not normalized, then it may occur that the trained classifier would be biased to certain features only and also training time increases and the accuracy decreases.

$N_2 = (N_1 * min) / (max-min)$

Where, N_2 = New Value

N_1 = Old Value

VI. FEATURE SELECTION TECHNIQUES

Feature Selection is the process of selecting a subset of relevant features for use in model construction. It improves the accuracy of model and also reduces over fitting.

A. Correlation-based Feature Selection

It is based on the assumption that features are conditionally independent given the class, where feature subsets are evaluated based on the following hypothesis: A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other. The evaluation function is described by following equation,

$$M_s = \frac{k\bar{r}_{cf}}{\sqrt{k+k(k-1)\overline{rff}}}$$

Where Ms is the heuristic "merit" of a feature subset S containing K features, $\overline{r_{cf}}$ is the mean feature-class correlation, and $\overline{r_{ff}}$ is the average feature-feature correlation.

CFS is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best first search and genetic search.

B. Entropy and Information Gain

Entropy is the average amount of information used to classify an object. It can be viewed as a measure of uncertainty of the system. Entropy of a discrete feature Y is defined as,

$$H(Y) = - \sum y \varepsilon Y \, P(Y) \log_2 (P(Y))$$

Information gain is used as a measure for evaluating the worth of an attribute based on the concept of entropy. Higher the entropy the more the information content. Information gain for two attributes X and Y is defined as,

$$IG(X, Y) = H(Y) – H(Y|X) = H(X) – H(X|Y)$$

From the above equation, Information gain is a symmetrical measure—that is, the amount of information gained about Y after observing X is equal to the amount of information gained about X after observing Y.

VII. CLASSIFIER

Classification is a data mining task that maps the data into predefined groups & classes. It is also called as supervised learning.

A. Random Forest

Random forest is an ensemble classifier. It has a higher classification accuracy compared to single decision tree. Random forest contains many decision trees which are trained with the same dataset and different features selected at random. It avoids over-fitting as features and data are randomly selected.

Steps for Random Forest Algorithm

1) Randomly select 'k' features from total 'm' features where k<<m.
2) Among the 'k' features, calculate the node'd' using the best split point.
3) Split the node into daughter nodes using the best split.
4) Repeat above steps until 'l' number of nodes has been reached.
5) Build forest by repeating above steps for 'n' number times to create 'n' number of trees.

B. Naïve Bayes

Naïve Bayes is a classification technique based on Bayes theorem with an assumption of independence among predictors. In simple terms, the presence of particular feature in a class is unrelated to the presence of any other feature. Bayes theorem is as follows:

$$P(c|x) = (P(x|c) * P(c)) / P(x)$$

$$P(c \mid x) = P(x_1|c) * P(x_2|c) * \ldots * P(x_n \mid c) * P(c)$$

C. J48

J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The algorithm generates the rules from which particular identity of that is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Basic Steps in the algorithm:

1) In case the instances belong to the same class the tree represents a leaf so that the leaf is returned by labelling with same class.
2) The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.
3) Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

VIII. CONCLUSION AND FUTURE SCOPE

We discussed about IDS which is capable of detecting both individual and combination attacks. Data pre-processing methods, feature selection techniques and various classifiers for NSL-KDD dataset was discussed briefly. The IDS proposed will be able to detect only specific attacks as we are using NSL-KDD. In the future, we can use more efficient and reliable machine learning algorithms for intrusion detection.

ACKNOWLEDGMENT

Our sincere thanks to S.R. Khonde for helping us in the development of this paper and inculcating her knowledge in this domain to us. Her constant motivation and support at every stage of development has helped us design this paper.

REFERENCES

[1] M.N. Napiah, M.Y.I. Idris, R. Ramli, I. Ahmedy , "Compression header analyzer Intrusion Detection System (CHA - IDS) for 6LoWPAN communication protocol" ,IEEE Access, Vol. **6, 2018**.

[2] P. Aggarwala, S.K. Sharma, "Analysis of KDD dataset attributes-class wise for intrusion detection", Procedia Computer Science, Vol. **57**, pp. **842–851, 2015**.

[3] H. Chae, B. Jo, S. Choi, T. Park, "Feature selection for intrusion detection using NSL-KDD", Recent Advances in Computer Science, pp. **184-187, 2013**.

[4] S.S. Panwar, Dr. Y. P. Raiwani, "Data reduction techniques to analyze NSL-KDD dataset", International Journal of Computer Engineering and Technology (IJCET), Vol. **5**, Issue **10**, pp**. 21-31, October (2014)**.

[5] H.P.S. Sasan and M. Sharma, "Intrusion detection using feature selection and machine learning algorithm with misuse detection", International Journal of Computer Science & Information Technology (IJCSIT), vol. **8**,no.**1**,pp. **17-25, 2016**.

[6] L.Dhanabal, Dr. S.P. Shantharajah. "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms" International Journal of Advanced Research in Computer and Communication Engineering, Vol. **4**, Issue **6, 2015**.

[7] A. Narayan and T.J. Parvat., "An Intrusion Detection System, (IDS) with Machine Learning (ML) model combining hybrid classifiers" Journal of Multidisciplinary

Engineering Science and Technology (JMEST), Vol. **2**, Issue **4, April - 2015**.

[8] P. Rutravigneshwaran, "A Study of Intrusion Detection System using Efficient Data Mining Techniques" International Journal of Scientific Research in Network Security and Communication, Vol. **5**, Issue **6, December – 2017.**

[9] M. Arora, S. Sharma, "Synthesis of Cryptography and Security Attacks" International Journal of Scientific Research in Network Security and Communication, Vol. **5**, Issue **5, October – 2017.**

[10] U. K. Singh, C. Joshi, S. K. Singh, "Zero day Attacks Defense Technique for Protecting System against Unknown Vulnerabilities" International Journal of Scientific Research in Computer Science and Engineering, Vol. **5**, Issue **1, February – 2017.**

[11] A. Ahmad , M. Asif, S. R. Ali, "Shallow Learning and Deep Learning Methods for Network security" International Journal of Scientific Research in Computer Science and Engineering, Vol. **6**, Issue **5, October – 2018.**

Authors Profile

P. Patil is a student currently studying in ModernEducation Society's College of Engineering, Pune. She is pursuing Bachelors in Computer Engineering in a 4 year program.

T. Bagwan is a student currently studying in Modern Education Society's College of Engineering, Pune. She is pursuing Bachelors in Computer Engineeringin a 4 year program.

S. Kulkarni is a student currently studying in Modern Education Society's College of Engineering, Pune. She is pursuing Bachelors in Computer Engineering in a 4 year program.

C. Lobo is a student currently studying in Modern Education Society's College of Engineering, Pune. She is pursuing Bachelors in Computer Engineering in a 4 year program.

S. R. Khonde is Assistant Professor in Modern Education Society's College of Engineering, Pune. She is currently Pursuing her Ph.D from Sathyabama Institute of Science and Technology, Chennai, India. She has published her works in 9 International Journals, 3 National Conferences and 3 International Conferences.