REVIEW OF VARIOUS INTRUSION DETECTION METHODS FOR TRAINING DATA SETS

Nilofer Shoaib Khan¹ and Prof. Umesh Lilhore²
¹ M.Tech Scholar NIIST Bhopal (MP)
² PG In charge NIIST Bhopal (MP)

Abstract- In the field of Information technology security plays a vital role. Unauthorized entries or any anomalies in system are known as intrusion and detection of these anomalies are known as Intrusion Detection System (IDS). As the attacks have increased in huge numbers over the past few years, IDS is increasingly becoming a critical component to secure the network. Designing of an efficient Intrusion detection system is always challenging tasks for the researchers. IDS performs monitoring and analyzing of network traffic for detecting security violations many researcher suggested data mining technique such as classification, clustering ,pattern matching and rule induction for developing an effective intrusion detection system. In this review paper we are presenting comparative analysis of various IDS by using decision tree based classifiers for data sets such ad Kyoto data sets, KDD-Cup 99 data sets. This study will help to develop an efficient IDS system for security system.

Keywords- Intrusion detection system, Intruder, Decision tree, and KDD-Cup 99, Kyoto datasets

I. INTRODUCTION

An Intrusion Detection System (IDS) is a system that monitors network to check harmful activities in the network and reports events that does not meet the security criteria to the network administrator. IDSs are categorized as Signature based and Anomaly based. Signature or Misuse based IDS uses various techniques to locate the similarity among system behavior and previously known attacks stored in the signature database. Anomaly based IDS detects activities in a network which deviates from normal behaviors stored in system profiles database. There are various classifiers that are applicable to misuse based detection. Some are tree based such as decision tree [1], and random forest [2], whereas some are rule based such as oneR [3], while some are function based such as SVM (Support Vector Machine) [4]. Anomaly detection is based on the prediction that normal user behavior is different from normal user. Using decision tree analysis, the logic of decision tree can be implemented in the intrusion detection system [7]. A decision tree is a method of predicting unknown information which are converted into a tree like structures[8]. Decision tree can prove itself more efficient in terms of retrieval of data for the purpose of making decisions. The decision tree starts with a root node which splits recursively per the possible conditions and its decision. In this paper various decision tree algorithms are described such as J48 and Id3 [1].

II. LITERATURE SURVEY

The Kyoto 2006+ [1] is an evaluation dataset of network detection mechanism obtained from diverse honey pots from November 2006 to August 2009. This dataset captures the real network traffic without any human alteration or deletion. It encompasses the recent trends of network attacks distinguished from normal traffic via the use of honey pots. It consists of 24 statistical features where 14
conventional features are extracted from KDDCUP'99 dataset [2], and 10 additional features are added that may enable to investigate more effectively what Kind of attacks happened in the networks.

The Kyoto dataset is labeled; the label indicates whether the session is an attack or not. In the original database, there are three labels: ‘1’ (normal session); ‘−1’ (known attack), and ‘−2’ (unknown attack). Nevertheless, since the unknown attacks in the database are extremely rare (0.7%), which makes them very difficult to detect for a machine learning model, we attribute a same label for known and unknown attacks, so that the problem becomes a binary classification [7].

The KDD Cup '99 dataset is the most well-known intrusion detection dataset available and researched by many researchers. The network traffic records in the dataset are classified as Normal or one of the four attack type’s i.e. DO - denial of service, PROBE- network probe, R2L- remote to local and U2R- user to root attacks. In past various static machine learning algorithms have been evaluated and results are published. The results of the KDD’99 classifier learning contest, as summarized by Elkan [3], were all variants of the C5 decision tree algorithm (see Quinlan [4]). After the contest a comprehensive set of other algorithms were tested on the KDD Cup 99 data, mostly with comparable results, were presented by Sabhani and Serpen [5], Sung and Mukkamala [6], Chavan, Shah et al. [7] and Peddabachigari, Abramham et al. [8]. The majority of results published are on the KDD Cup '99 `10%' training set only see Sung and Mukkamala [6], Kayacik, Zincir-Heywood et al. [9] and Lee, Shin et al. [10]. Some of the researchers extracted 11,982 records from KDD Cup 10% training dataset and build custom training datasets with 5,092 records and 6,890 test record see Chavan, Shah et al. [7], Chebrolu, Abraham et al. [11] and Chen, Abraham et al. [12]. Chavan, Shah et al. [7] use a decision tree method for ranking of features per class. They reduced number of features from 41 to 13 for 'normal',16 for `probe', 14 for `dos', 15 for `u2r' and 17 for `r2l' for experiment they evaluated it using artificial neural networks and fuzzy inference systems. Kayacik, Zincir-Heywood et al. [9] investigated the relevance of each feature provided by the KDD Cup '99 intrusion detection dataset in terms of information gain and presented the most relevant feature for each individual attack. Another important result was that 9 features do not make any contribution for intrusion detection.

Tavallaee and Bagheri et al. [15] described the importance of each feature in KDD '99 intrusion detection dataset for detection of DOS, PROBE, U2R L2R and Normal class. They also discuss various problems of KDD Cup ‘99 datasets and created a revised version of the datasets, called NSL-KDD to address the some of known issues. They modified the class distributions by cleaning the training and testing datasets. This will avoid biasness towards the more frequent records.

Ben Amor et al. [16] performed comparative analysis of decision tree vs naïve bayes and found that decision tree performs slightly better than naïve bayes. They also found that building naïve bayes computational model is faster than of decision tree. Decision trees generally have very high speed of operation and high attack detection accuracy.

### III. INTRUSION DETECTION SYSTEM

Intrusion-detection-system recognized the known and the unknown patterns of the attacks over the network after which this system performs the required actions according to the detected intrusion for the connections of network. Intrusion is defined as the any type of group of the actions which are trying to affect the confidentiality, integrity or the availability of the system and the intrusion-detection-system is the device or the software application which monitors the traffic of network for detecting the suspicious activity, if any activity of this type found then an alert is generated for the system or the network administrator in order to warn them [5,7]. The major three modules of the intrusion-detection system are-

- **System Monitoring**
- **Activity Analyses**
- **Action Phase or Response**

  The Intrusion-Detection-Systems is capable of finding the attacks within the existing environments and it resolve the uncertainties within the monitoring of network through implementing the detection systems over the area of attacks [13]. The intrusion-detection-system is capable to provide the prevention commands also to the firewalls and the access control that changes to the routers which may be considered as an enhancement over the technologies of firewall. This may provide the decisions for access control which are dependent on the application content, no on the IP address or the ports similar to the basic types of firewalls [1, 9].

3.1 **Types of IDS** -Here are described some types of Intrusion-detection-systems, which are of three types mainly termed as-

- **Host based IDS or HIDS**
- **Network based IDS or NIDS**
- **Distributed IDS or DIDS**

A. **HIDS**- This type of intrusion-detection-system monitors the information and then analyse it which is gathered from the particular host system. The host-based IDS are executing over the host machine that detects the intrusion through gathering the information like the file system which is used, the events on the network and the system calls etc for detecting the intrusion. This type of IDS analyse the modification found within the host kernel, in the behavior of program and with the host file system [4].

B. **NIDS**-This type of IDS is attempt to find the malignant activity over the network like DoS attacks, scan the ports or may try to crack the computers through monitoring the network traffic even. And the collected information from the network is then got compared with the known patterns of intrusion for the detection of intrusion [11].

C. **DIDS**-This type of IDS includes various types of other IDS like Network based IDS, Host based IDS etc, that are found over the huge network, and these all may interact with each other, or through the central server which provides the monitoring of network.

3.2 **Intrusion Detection Methods**-

There are various techniques that have been applied for the Intrusion-detection-system in order to detect the intrusion and provide the security to the system. Here are described some of the techniques of the IDS that provides the security in some way-

A. **Statistical Techniques**- In this technique, dependent on some particular conditions the statistical comparison have made on the data which is gathered from the network or the system in order to find the intrusion.

B. **Pattern recognition Techniques**-Within this technique, the major advantage is by applying the sequence of the penetration scenarios which are embedded within the system, decreases the requirement to analyse the huge amount the data.

C. **Rule based Techniques**-In this the set of if-then1 rules are implemented to detect the attacks in which every rule is linked with a particular operation within the system.

3.3 **Model Classifications**- Following classification are use

3.3.1 **J-48 Decision Trees**-

It is a Java implementation of the C4.5 algorithm in the weka which is an open source data mining tool [4]. It builds decision trees from a set of training data, using information entropy. At each node of the tree, J48 chooses one attribute of the data that most effectively splits its set of samples into subsets. Its criterion is the highest and normalized information gain that results from choosing an attribute for splitting the data. The attribute which has highest normalized information gain is chosen to
make the decision. For each attribute, the gain is calculated and the highest gain is used in the decision node.

3.3.2 ID-3 Decision Trees-ID-3 or Iterative Dichotomiser 3 Algorithm [5] is a Decision Tree algorithm. It builds the tree from the top to down, with no back tracking. It makes use of Information Gain to select the most useful attribute for classification. ID3 is based on the Concept Learning System (CLS) algorithm. It is a Class for constructing an unpruned decision tree. ID3 has a drawback that it can only deal with nominal attributes. No missing values allowed in ID3. Empty leaves may result in unclassified instances [15].

3.3.3 Simple K-Mean Clustering- K-mean algorithm is a fast clustering algorithm to divide the data into k groups [8]. First, it selects k points as the centroid of the k clusters. Then it calculates the Euclidean distance of each data point to centroid of each cluster. Data point is assigned to the cluster which has minimum Euclidean distance [3]. New centroids are calculated by evaluating the mean of each cluster data points. The process is repeated until specified iterations achieved or same centroid evaluated in successive iterations.

3.4 Weka Tool-WEKA is a data mining system developed by the University of Waikato in New Zealand that implements data mining algorithms. WEKA is a state-of-the-art facility for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tools. The new machine learning schemes can also be developed with this package. WEKA is open source software issued under the GNU General Public License [13].

Advantages of Weka include-
- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.

IV. PROBLEM IDENTIFICATION

There are many existing mechanisms for Intrusion detection system, but the major issue is the security and accuracy of the system. To improve the problem of accuracy and the efficiency of the system a very common classification approach i.e. decision tree is used.

In this paper, we present a new evaluation dataset, called Kyoto 2006+, built on the 3 years of real traffic data (Nov. 2006 ~ Aug. 2009) which are obtained from diverse types of honey-pots. Kyoto 2006+ dataset will greatly contribute to IDS researchers in obtaining more practical, useful and accurate evaluation results. Furthermore, we provide detailed analysis of honey-pot data and share our experiences so that security researchers are able to get insights into the trends of latest cyber attacks and the Internet situations.

- Attacks
  - Can not reflect current malicious activities
  - Stealthy scan ⇒ short time interval, no multiple IP address scan
  - No attacks against Windows machines

- Protocol types
  - Only TCP, UDP, ICMP
  - Cannot detect attacks such as ARP Spoofing
Simplicity
- Only 3 real victim hosts
- 1000’s of virtual hosts and 100’s of user automata(custom software)

V. OBJECTIVE OF THE WORK

Main objective of the work is to presents detailed study of IDS system and data sets such as KDD Cup-99 and Kyoto data sets. This study will helps us to develop more efficient and accurate IDS detection method. Proposed method wills overcome problems which are discussed in section 4.

VI. CONCLUSION AND FUTURE WORK

In the research paper presents a novel study of IDS system. Intrusion detection is a challenging task. We have also discussed Kyoto 2006+ dataset built on 3 years of honey pot data. It consists of 14 statistical features derived from KDD Cup 99 dataset as well as 10 additional features which can be used for further analysis and evaluation of NIDSs. By using Kyoto 2006+ dataset, IDS researchers are able to obtain more practical, useful and accurate evaluation results.

In future work we will developed an efficient network based IDS detection method for Kyoto 2006+ data sets by using Weka tool and will compare its results with various existing classification methods.

REFERENCE

[1] Maryam M. Najafabadi, Taghi M. Khoshgoftaar and Naeem Seliya, Evaluating Feature Selection Methods for Network Intrusion Detection with Kyoto Data, International Journal of Reliability, Quality and Safety Engineering, World Scientific, Vol. 23, No. 1 (2016), PP 1650001-1650022
[2] Jamal Hussain, Samuel Lalmanawma, Feature analysis, evaluation and comparisons of classification algorithms based on noisy intrusion dataset. ScienceDirect,2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016), PP 188 – 198
[3] Safaa O. Al-mamory , Firas S. Jassim, On the designing of two grains levels network intrusion detection system, ScienceDirect, Karbala International Journal of Modern Science 1 (2015) 15e25
[4] G.V. Nadiammai, M. Hemalatha, Effective approach toward Intrusion Detection System Using data mining techniques, Egyptian Informatics Journal, Elsevier, 2013, PP 37-50.
[5] Ghorbani AA, Lu W, Tavallaee M. Network Intrusion Detection and Prevention Concepts and Techniques, Springer New York Dordrecht Heidelberg London; 2010.
[6] Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications 2014; 41(4): 1690–1700
[7] Shwarz SS, Singer Y, Srebro N. Pegasos: Primal and gradient solver for SVM. In: Proceedings of the 24th International conference on machine learning; 2007. p. 807-814.
[8] Freund Y, Schapire RE. Large margin classification using the perceptron algorithm. Machine Learning 1999; 37(3): 277-296.
[9] Khardon R, Wachman G. Noise tolerant variants of the perceptron algorithm. The Journal of Machine Learning Research 2007; 8: 227-248.
[10] Wettschereck D, Dietterich T. Improving the performance of radial basis function networks by learning center locations. In Neural Information Processing Systems 4. Denver, CO: Morgan Kaufmann NIPS 1992; 4: 1133–1140.
[11] Dong L, Frank E, Kramer S. Ensembles of balanced nested dichotomies for multiclass problems. In: Proceedings of the 9th European conference on Principles and practice of knowledge:2005.
[12] Bottou L. Online Algorithms and Stochastic Approximations. Online Learning and Neural Networks.Cambridge University Press; 1998.
[13] Cohen WW. Fast effective rule induction. In Proceedings of the 12th International Conference on Machine learning; 1995. p. 115–123.
[14] Breiman L. Random Forests. Machine Learning 2001; 45(1): 5-32.
[15] Kohavi R. The power of decision tables. In: Proceedings of the 8th European Conference on Machine Learning; 1995. p. 174-189.
[16] Kohavi R. Scaling up the accuracy of Naive Bayes Classifier: a Decision-Tree Hybrid. In: Proceedings of the 2nd International conference on knowledge discovery and data mining; 1996. p. 202-207.
[17] Howlett RJ, Lakhmi CJ. Radial Basis Function Networks 2: New Advances in Design. Springer-Verlag Berlin Heidelberg; 2001.

[18] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter 2009; 11(1): 10–18.