Volume 6, Issue 8, August 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey of Attack and Weka Tool Using IDS Technique

Baljeet*
Research Scholar, Department of Computer Sc & Engg.,
Kurukshetra University, Haryana, India

Parbhat Verma
Asst. Prof., MIET, Department of Computer Sc & Engg,
Kurukshetra University, Haryana, India

Abstract— Intrusions detections systems from point of view of security policy are a second line of defense; they have a supervisory role to observe the activities of our network or hosts to identify attacks in real time. In our days, electronics attacks can cause a very destructive damage for nations which make necessary the use of completed security policy to minimize the potential threats. IDS it is a very important element to resist against this vulnerability,. we study a wired data base Knowledge Discovery Data Mining (KDD) CUP 99 and a Data Mining Tools Waikato Environment for Knowledge Analysis (WEKA) to combine the advantages of an intrusion detection algorithm (PART).

Keywords— IDS, Testing, Redundant, classification

I. INTDOUCTION

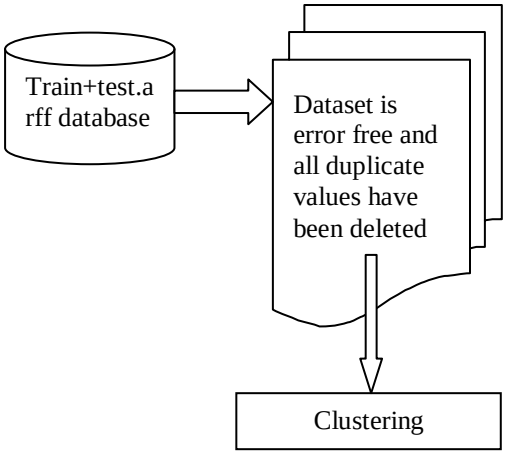
Intrusion Detection Systems (IDS) are now becoming one of the essential components in any organization'snetwork. IDS are designed to detect any intrusion or hostile traffic in a network. With the serious need of such detection systems organizations have been investing to produce a more effective IDS. Intrusion Detection Systems can be implemented as a hardware based or software-based [2,3]. The later type of IDS is more configurable and easy to update while the hardware based is designed to handle large amount of trafficbut more expensive and require more maintenance. There is therefore a need to evaluate the available software-based IDS[4].

1.1 Testing & Training Data

In this study we are taking dataset from the UC [knowledge Discovery Database Archive. KDD'99 dataset. The technique of monitoring and keeping secure systems, it is Very important to test and train intrusion system using a huge amount of intrusion data. There could be two possibilities of capturing data. 1. Scanning the port. ii. Or to use the dummy data from KDD. KDD process is interactive and iterative dataset. The data set is a huge data which is in the .txt format and .arff format (attribute relation file format) [5].

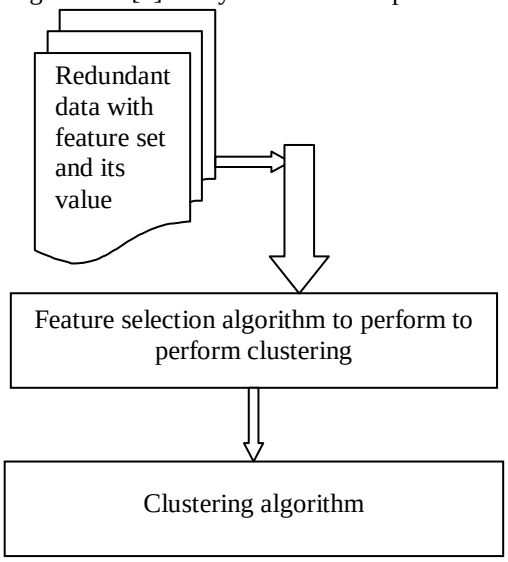
1.2 Redundant Data

Test data is rectified through feature selection algorithm, by this algorithm dataset become redundant so that algorithm complexity could be reduced. To handle with data mining or data ware house, first we have to deal with detecting and removing duplicate data from database. This step is the most important and crucial to be performed. Dealing with these steps can make our results more appropriate and performance more efficient. It has been found that the similar logical data has multiple representations in the database. Removing multiple representations is difficult because it causes different types of errors. It is not important to just cleaning the multiple representation but also verifying that data should be authenticated and should be real-world entity, else it would be [5] far away from the results expected. In this research process it is meant to be initial step for processing so that algorithms applied to this redundant data is real-world entity. After this process clusters are formed by using feature selection techniques so that we can later use them for rule generation and for determine the nature of attacks in the transactions.



1.3 Classification Of Data

In this study we are using a hybrid algorithm in contrast with K means and NaiVe base algorithm. K means is a Classification algorithm. It basically divides each transaction into clusters so that they can be easily distinguished according to the particular features representation. These steps are particular to the algorithm going to use. Classifications help in defines clusters into specific classes of features. Although these steps don't actually form clusters or classifies data set here at the processing of specific algorithms [5]. They are the end steps to define the rule generations



In order to evaluate the performance of classification techniques, an overall methodology consisting of four steps, i.e., dataset preparation, data pre–processing, attribute selection, and classification, is proposed [6].

II. DATASET PREPARATION

To standardize the dataset, KDD CUP 1999 [9] was selected for our classification evaluation process. In general, KDD CUP 1999 is based on the intrusion detection simulation of U.S. Air force local area networks via *tcpdump* [www.tcpdump.org]. The dataset consists of network access behavior including up to 41 attributes as well as heterogeneous access patterns. In general, KDD CUP consists of four main attacks; namely, DoS (Denial of Service), PROBE, U2R (User to Root), and R2L (Remote to User) excluding BOTNET attacks, each of which generates each individual attack class as shown in Table I and II. In addition to a traditional KDD CUP 1999, in this research, an extra type of recent attack was evaluated, HTTP BOTNET. Here, we emulated the Zeus attack (version 1 and 3) via *tcpdump* output acquiring from Zeus traces [10]; and then we ordered the packet sequence numbers for each connection, i.e., from SYN to FIN packets to suitably match KDD CUP 1999 formats [6].

- DoS: This attack can freeze the server operation and activity by acquiring all resources so that the server cannot provide any service, commonly using flooding–based schemes.
- PROBE: This attack is used during a preparation stage for other attacks in order to gain valuable information such as enabled ports and services as well as Internet address information.
- U2R: This attack performs a specific operation in order to penetrate into a system hole/leak such as Buffer Overflow.
- R2L: The attack is used to take advantages of related users' safety information or configuration such as SQL Injection.
- BOTNET: This attack is to run the computer into a bot (zombie) to perform a particular task over the Internet as administrator.

2.1 Data Pre-Processing

Before performing data mining classification using Weka tools, a traditional KDD CUP 1999 is required to transform into a suitable format, i.e., *.csv*, *.arff*, and *.cs45*, and here *.arrf* was chosen. Extra information is also required as examples below. - @relation *name*: to indicate the dataset unique name. - @attribute *at–name type*: to indicate characteristic of attribute such as "duration" with "real" and "src_byte" with "numeric". - @data: to indicate the end of header. In addition, to make evaluated parameters fairly affected by un–equivalent numbers of records, a random record selection was performed given an equal proportional number of evaluated classes [5].

2.2 Attribute Selection

Since there are too many un–relevant attributes probably leading to low classification precision and high computational complexity, this selection stage is used to figure out the suitable attributes applying the information gain to segregate the dataset in that each gain will be computed for each data dimension, and then the highest gain will be selected as the evaluated record representation stated in equations 1 and 2 resulting in the attribute lists shown in Table III. Here, *a* is a member of the set of attributes. *S* denotes as the set of all training examples, and *p* is the probability of *s* in the set *S*. It should be noted that each of the four classes was selected based on the number of appropriate records for

performance evaluation purposes; however, to compute information gain [5], the larger the records, the higher the accuracy observed during our intensive empirical evaluation results.

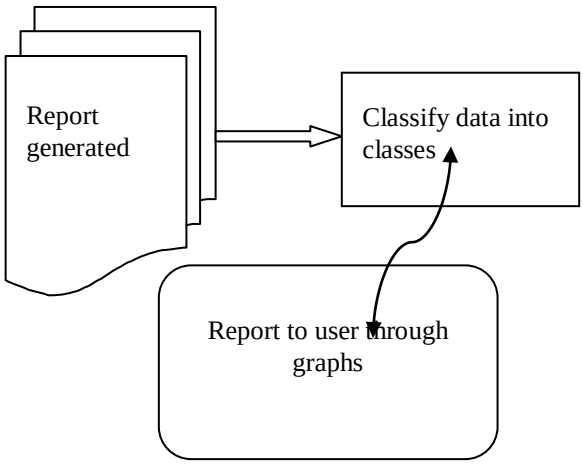
III. CLASSIFICATION

There are six main classification models embedded into the recent Weka tools; namely, Decision Tree, Ripper Rule, Neural Networks, Naïve Bayes, *k*–Nearest– Neighbor, and Support Vector Machine [12–8].

1) **Decision Tree (J48) is** one of the tree classifications Techniques in which a particular tree will be generated given nodes as attributes, leafs as classes, and edges as testing results.
2) **Ripper Rule (JRIP)** is used to generate various rules by adding repetitive datasets until the rules cover all data pattern according to the training dataset. In addition, once all rules are generated, some of them will be merged in order to reduce size.
3) **Neural Networks (MLP – Multilevel Perceptron**) has a distinctive characteristic as a three layered feed--- forward neural network: one input, one hidden, and one output layer. In order to link each node in each level, it may include an additional weight to properly adjust the path traversal selection process.
4) **Naïve Bayes is derived from Bayes'** Theorem by applying the learning probabilistic knowledge for classification given the assumption that the predictive attribute is conditionally independent based on each individual class.
5) ***k*–Nearest–Neighbour (IB**K) is used to perform the classification considering *k* sub–datasets, each of them has similar characteristics applying Euclidean Distance to figure out the group, and here, IBK is one of the simplified *k*–Nearest–Neighbour classifiers. .
6) **Support Vector Machine (SMO)** is basically a linear classifier (two–classes) used to figure out the largest distance between two sets, and SMO is the sequential minimal optimization algorithm for training SVM using polynomial or Gaussian kernels.[5]

IV. RULE MINIMG

It is a type of module where rules will be generated according to the report produces after applying association rule algorithm. [N this workflow user will be alerted by graphical representation and analysis of the report generated by the hybrid algorithm. This will provide all the attacks occurred in the database.



V. CONCLUSION

WEKA is a powerful instrument that offers several data Pre-processing facilities as well as facilities for their Analysis through classification, regression, clustering, Association rules techniques, etc. WEKA's connection to different databases, although possible, is difficult andRequires knowledge of the SQL language, which limits the number of WEKA users. The contribution brought by this consists of an original solution that greatly facilitates the loading process of the databases in WEKA.

REFERENCES

[1] Mouaad KEZIH, Mahmoud TAIBI**" Evaluation Effectiveness of Intrusion Detection System with Reduced Dimension Using Data Mining Classification Tools",** 2013.
[2] Wallner, R., **Intrusion Detection Systems**. 2007.
[3] Di Pietro, R. and L.V. Mancini, **Intrusion detection systems. 2008:** Springer Verlag.
[4] Adeeb Alhomoud, Rashid Munir" **Performance Evaluation Study of Intrusion Detection Systems**" The 2nd International Conference on Ambient Systems,
[5] Chakchai So–In, Nutakarn Mongkonchai, Phet Aimtongkham. *"***An Evaluation of Data Mining Classification Models for Network Intrusion Detection"**IEEE 2014.
[6] T.R. Gopalakrishnan Nair, K.Lakshmi Madhuri" **Data Mining USsing Hierachical Virtual K-Means Apporaoch Intergrating Data Fragments In Cloud Computing Envioroment "**IEEE 2011.
[7] Kapil Wankhade, Sadia Patka **" An Efficient Approach for Intrusion Detection Using Data Mining Methods"**IEEE 2013.

[8] Chirag N. Modi, Dhiren R. Patela, Avi Patelb, Muttukrishnan Rajaraja **" Integrating Signature Apriori based Network Intrusion DetectionSystem (NIDS) in Cloud Computing**"2012.

[9] S.V. Shirbhate, Dr.S.S.Sherkar ,Dr.V.M.Thakare " **Performance Evaluation of PCA Filter In Clustered Based Intrusion DetectionSystem"**2014.

[10] Cheung-Leung Lui ,Tak-Chung Fu, "**Agent-based Network Intrusion Detection System Using Data Mining Approaches"**IEEE 2005.

[11] Kailas Elekar, Amrit Priyadarshi M.M. Waghmare " **Use of rule base data mining algorithm for Intrusion Detection"** IEEE 2015. International Conference on Pervasive Computing (ICPC) -2015 IEEE

[12] Mohammed M. Alani" **MANET Security: A Survey"** 2014 IEEE International Conference on Control System, Computing and Engineering, 28 - 30 November 2014, .

