International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 018-021
e-ISSN:2278-621X

AN ANALYSIS OF INTRUSION DETECTION SYSTEMS USING KDD DATASET IN WEKA

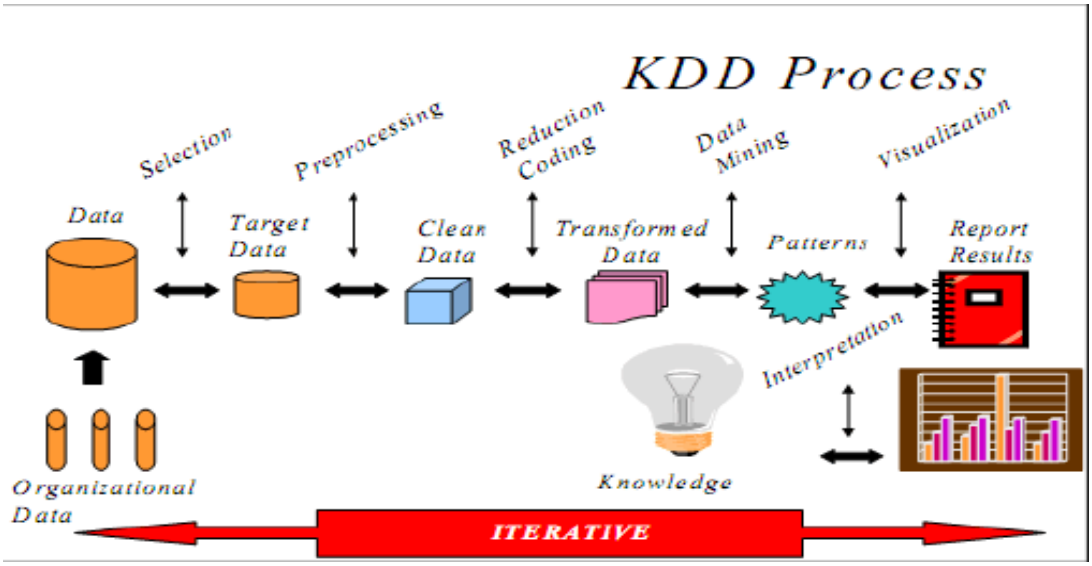
L.Sheeba[1]

Abstract-This paper presents the analysis of the effect of clustering the training data and test data in classification efficiency of Naive Bayes classifier. KDD cup 99 benchmark dataset is used in this research. Intrusion Detection System (IDS) is a security system that acts as a protection layer to the infrastructure. But selecting important features from input data lead to a simplification of the problem, faster and more accurate detection rates. The relevance of each feature in KDD '99 intrusion detection dataset to the detection of each class. The IDS for detecting the attacks effectively has been proposed and implemented. For this purpose, a new feature selection algorithm called Optimal Feature Selection algorithm based on Information Gain Ratio has been proposed and implemented. This feature selection algorithm selects optimal number of features from KDD Cup dataset. In addition, two classification techniques namely Support Vector Machine and Rule Based Classification have been used for effective classification of the data set. This system is very efficient in detecting DOS attacks and effectively reduces the false alarm rate. The proposed feature selection and classification algorithms enhance the performance of the IDS in detecting the attacks.
Keywords – Support Vector Machine, KDD, DOS.

1. INTRODUCTION

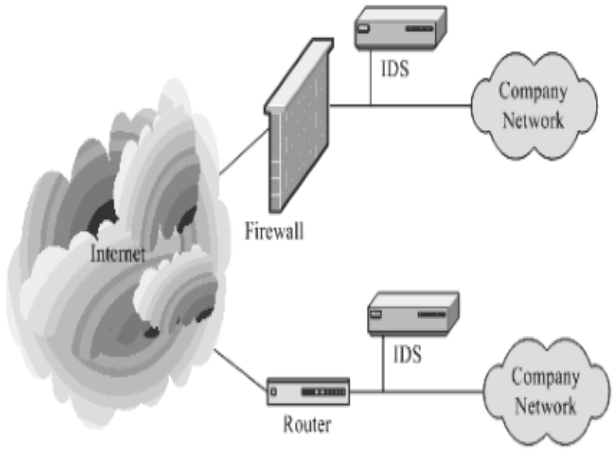
Intrusion detection begins where the firewall ends. Preventing unauthorized entry is best, but not always possible. It is important that the system is reliable and accurate and secure. Intrusion detection is defined as real-time monitoring and analysis of network activity and data for potential Vulnerabilities and attacks in progress. One major limitation of current intrusion detection system (IDS) technologies is the requirement to filter false alarms. IDS is defined as a system that tries to detect and alert of attempted intrusions into a system or a network[1].IDSs are classified into two major approaches. Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices. Intrusion prevention is the process of performing intrusion detection and attempting to stop detected possible incidents. Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, attempting to stop them, and reporting them to security administrators. In addition, organizations use IDPSs for other purposes, such as identifying problems with security policies, documenting existing threats and deterring individuals from violating security policies. IDPSs have become a necessary addition to the security infrastructure of nearly every organization. IDPSs typically record information related to observed events, notify security administrators of important observed events, and produce reports. Many IDPSs can also respond to a detected threat by attempting to prevent it from succeeding. They use several response techniques, which involve the IDPS stopping the attack itself, changing the security environment (e.g., reconfiguring a firewall), or changing the attack's content.

[1] Department of BCA, Assistant professor,PSGR Krishnammal College for Women, Coimbatore Tamilnadu, India



2. TYPES OF IDS'S

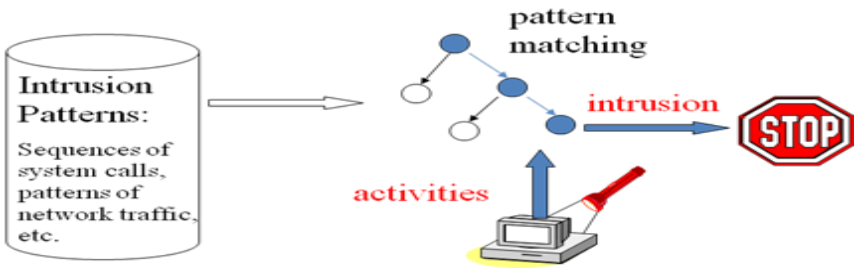
Several types of IDS technologies exist due to the variance of network configurations. Each type has advantages and disadvantage in detection, configuration, and cost.



- Signature based detection.
- Anomaly based detection.

2.1 Signature based detection:

Signature based IDS are based on looking for "known patterns" of detrimental activity. The Benefits are Low alarm a rate of all it has to do is to look up the list of known signatures of attacks and if it finds a match report it. Signature based NID are very accurate. Speed the systems are fast since they are only doing a comparison between what they are seeing and a predetermined rule. Intrusion Detection system is programmed to interpret a certain series of packets, or a certain piece of data contained in those packets, as an attack. For example, an IDS that watches web servers might be programmed to look for the string "phf" as an indicator of a CGI program attack. Most signature analysis systems are based off of simple pattern matching algorithms. In most cases, the IDS simply looks for a sub string within a stream of data carried by network packets. When it finds this sub string (for example, the ``phf'' in ``GET /cgi-bin/phf?''), it identifies those network packets as vehicles of an attack.



2.2 Anomaly based detection:

Anomaly based IDS are based on tracking unknown unique behavior pattern of detrimental activity. An IDS that looks at network traffic and detects data that is incorrect, not valid, or generally abnormal is called anomaly based detection. This method is useful for detecting unwanted traffic that is not specifically known. For instance, anomaly based IDS will detect that an Internet protocol (IP) packet is malformed. It does not detect that it is malformed in a specific way, but indicates that it is anomalous.

3. EXPERIMENTAL SETUP

3.1 Weka (Waika To Environment For Knowledge Analysis)

WEKA is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. *Weka* Waikato Environment for Knowledge Analysis (Weka)15 is a data mining tool available free of cost under the GNU General Public License. The version used in this study is 3.7.11 that has many state of the art machine learning tools and algorithms for data analysis and predictive modeling. This tool accepts the data file either in comma separated value (csv) or attribute-relation file format (arff) file format. For the simulation, arff files is already available with 42 attributes whereas arff files with lesser attributes as discussed in research methodology section are created through the pre-processing tab of the tool.

The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.

3.2 Kdd cup 99 Dataset Description

The steps followed as part of the research methodology are as follows:

KDD data set is selected10

Weka Tool is chosen for simulation

Random Tree is used as a binary classifier for simulation on Weka classifies the instances as attack or normal Preprocessing of training and testing data file with 42 attributes is done to generate 14 new training data files for each combination as discussed in Table 2

Every pair of 15 data set files (training and test), is simulated on random tree algorithm and the results are tabulated in Table 6.It must be noted that all 15 training and test data files as in Table 2, the last attribute of the original data set, that is, 'class' attribute is included.

3.3 Metrics

Intrusion detection metrics helps evaluate the performance of an intrusion detection system21. Some of the commonly used evaluation metrics used with respect to intrusion detection are False Alarm Rate (FAR), Detection Rate (DR), Accuracy, Precision, Specificity, F-score13.

All these evaluation metrics are basically derived from the four basic attributes of the confusion matrix depicting the actual and predicted classes. These elements of the confusion matrix are:

True Negative (TN): Number of instances correctly predicted as non-attacks.

False Negative (FN): Number of instances wrongly predicted as non-attacks.

False Positive (FP): Number of instances wrongly predicted as attacks.

True Positive (TP): Number of instances correctly predicted as attacks.

As shown in the Table 6, all the metrics are generated from these four basic elements. In this paper, two of the evaluation metrics that are considered for this study are FAR which is defined as the rate at which normal instances are classified as anomalous and DR which is defined as the ratio of number of instances of correctly predicted attacks to the total number of actual attack instances. Another metric used for the analysis of results is the graphical plot known as the Receiver Operating Characteristic (ROC) curve. It is a plot of DR and FAR. Though this curve does not exactly tell the best classification results in terms of FAR and DR but the area under the ROC curve helps to find Accuracy, performance, false positive rate.

4. CONCLUSION

In this paper, we statistically analyzed the entire KDD data set. The analysis showed that there are two important issues in the data set which highly affects the performance of evaluated systems, and results in a very poor evaluation of anomaly detection approaches. To solve these issues, we have proposed a new data set, NSL-KDD, which consists of selected records of the complete KDD data set. This data set is publicly available for researchers through our website and has the following advantages over the original KDD data set: It does not include redundant records in the train set, so the classifiers will not be biased towards more frequent records. There is no duplicate record in the proposed test sets; therefore, the performance of the learners is not biased by the methods which have better detection rates on the frequent records. The number of selected records from each difficulty level group is inversely proportional to the percentage of records in the original KDD data set.

5. REFERENCES

[1] Bace, R. and Mell, P. (2001). Intrusion Detection System, NIST Special Publications SP 800. November.

[2] Ajith, A., Ravi J., Johnson T. and Sang, Y.H. (2005). D-SCIDS: Distributed soft computing intrusion detection system, Journal of Network and Computer Applications, Elsevier, pp. 1-19.

[3] P. Garcı´a-Teodoro, J. Dı´az-Verdejo, G. Macia´-Ferna´ndez, E. Va´zquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges", computers & security, Vol. 28, pp. 18-28, 2012.

[4] J. Zhang and M. Zulkernine, "Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection", Symposium on Network Security and Information Assuranc-Proc. of the IEEE International Conference on Communications (ICC), Istanbul, Turkey, **(2006)** June.

[5] M. Tavallaee, E. Bagheri, W. Lu and A. Ghorbani, "A Detailed Analysis of the KDD'99 CUP Data Set", The 2nd IEEE Symposium on Computational Intelligence Conference for Security and Defense Applications (CISDA), **(2009)**.

[6] S. Mukkamala, G. Janoski, A. Sung, "Intrusion detection using neural networks and support vector machines" Proceedings of the 2002 IEEE International Joint Conference on Neural Networks, pp. 1702 – 1707.

[7] M. Sabhnani and G. Serpen, "Application of Machine Learning Algorithms to KDD Intrusion Detection Dataset within Misuse Detection Context", The 2003 International Conference on Machine Learning; Models, Technologies and Applications, pp. 209-215.

[8] J. B. D. Cabrera, C. Gutiérrez, and R. K. Mehra, "Ensemble methods for anomaly detection and distributed intrusion detection in Mobile Ad-Hoc Networks," Inf. Fusion, vol. 9, no. 1, pp. 96 – 119, 2008.

[9] S. T. Powers and J. He, "A hybrid artificial immune system and Self Organising Map for network intrusion detection," Inf. Sci., vol. 178, no. 15, pp. 3024 – 3042, 2008.

[10] F. Amiri, M. R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," J. Netw. Comput. Appl., vol. 34, no. 4, pp. 1184 – 1199, 2011.

[11] Y. Wei and M. Wu, "KFDA and clustering based multiclass SVM for intrusion detection," J. China Univ. Posts Telecommun., vol. 15, no. 1, pp. 123 – 128, 2008.

[12] X. Gan, J. Duanmu, J. Wang, and W. Cong, "Anomaly intrusion detection based on PLS feature extraction and core vector machine,"Knowl.-Based Syst., vol. 40, pp. 1 – 6, 2013.