A Review on Intrusion Detection System on KDDCUPS'99 Dataset with the Help of Different Techniques

Preeti Yadav
M Tech,CSE, Department of Computer Science & Engineering,
Guru Kashi University,Talwandi Sabo Bathinda Punjab.

Jaspreet Kaur
Assistant Professor,CSE, Department of Computer Science & Engineering,
Guru Kashi University,Talwandi Sabo Bathinda Punjab.

Abstract – Intrusion Detection System (IDS) becomes a crucial part of each laptop or network system. Intrusion detection (ID) may be a mechanism that has security for each computers and networks. Feature choice and have reduction is very important space of analysis in intrusion detection system. The dimensions and attribute of intrusion file are terribly massive. Within the analysis work intrusion detection totally different issues are two-faced. A hybrid model for feature choice and intrusion detection is very important issue in intrusion detection. The choice of feature in attack attribute and traditional traffic attribute is difficult task. There are choice of better-known and unknown attack is additionally two-faced a tangle of classification. There's multiclass downside throughout the classification of knowledge. Intrusion detection may be a downside of transportation infrastructure protection attributable to the actual fact that laptop networks are at the core of the operational management of a lot of of the nation's transportation. During this work intrusion detection is enforced on KDDCups'99 Dataset.

Index Terms – Intrusion, SVM, PCNN, KDDCups'99.

1. INTRODUCTION

During the last few years there is a dramatic increase in growth of computer networks. There are various private as well as government organizations that store valuable data over the network. This tremendous growth has posed challenging issues in network and information security, and detection of security threats, commonly referred to as intrusion, has become a very important and critical issue in network, data and information security. The security attacks can cause severe disruption to data and networks. Therefore, Intrusion Detection System (IDS) becomes an important part of every computer or network system. Intrusion detection (ID) is a mechanism that provides security for both computers and networks. Feature selection and feature reduction is important area of research in intrusion detection system. The size and attribute of intrusion file are very large. Due to large size of attribute the detection and classification mechanism of intrusion detection technique are compromised in terms of detection rate and alarm generation. For the improvement of intrusion detection process various authors and researchers work together for feature reduction and feature selection for intrusion detection system. In current scenario the feature reduction and selection process focus on entropy based technique. Some authors used neural network model such SOM and RBF neural network model for classification of intrusion data during attacking mode and normal mode of network traffic. On the mechanism of detection intrusion detection divide into two section host based intrusion detection system and network based intrusion detection system. Host based intrusion detection system in generally known as signature based intrusion detection system. Instead signature based intrusion detection system come along with another variant is called anomaly based intrusion detection. In anomaly based intrusion detection various technique are used such as supervised learning and unsupervised learning. In network intrusion Detection, independent and redundancy attributes leads to low detecting rate and speed of classification algorithms. Therefore, how to reduce network attributes to raise performance of classification algorithms by applying optimal algorithm has become a research branch of intrusion Detection. A new approach for network intrusion detection feature selection based on PCNN-SVM attributes selection and reduction. The available approaches for intrusion detection focus on improving detection accuracy and restraining false alarms, and given enough time, most of them can achieve satisfactory results in terms of these criteria. However, in practice, intrusion detection is a real-time critical mission, that is, intrusions should be detected as soon as possible or at least before the attack eventually succeeds.

2. INTRUSION DETECTION SYSTEM

An intrusion is an attempt to compromise the integrity, confidentiality, availability of a resource, or to bypass the security mechanisms of a computer system or network. James

Anderson introduced the concept of intrusion detection in 1980.It monitors computer or network traffic and identify malicious activities that alerts the system or network administrator against malicious attacks. Dorothy Denning proposed several models for IDS in 1987. Approaches of IDS based on detection are anomaly based and misuse based intrusion detection. In anomaly based intrusion detection approach.The system first learns the normal behavior or activity of the system or network to detect the intrusion. If the system deviates from its normal behavior then an alarm is produced. In misuse based intrusion detection approach IDS monitors packets in the network and compares with stored attack patterns known as signatures. The main drawback is that there will be difference between the new threat discovered and signature being used in IDS for detecting the threat. Approaches of IDS based on location of monitoring are Network based intrusion detection system (NIDS) and Host-based intrusion detection system (HIDS). NIDS detects intrusion by monitoring network traffic in terms of IP packet. HIDS are installed locally on host machines and detects intrusions by examining system calls, application logs, file system modification and other host activities made by each user on a particular machine.

3. FEATURE SELECTION

Due to the large amount of data flowing over the network real time intrusion detection is almost impossible. Feature selection can reduce the computation time and model complexity. Research on feature selection started in early 60s. Basically feature selection is a technique of selecting a subset of relevant/important features by removing most irrelevant and redundant features from the data for building an effective and efficient learning model.

4. DATA MINING, KDD, AND RELATED FIELDS

The term data mining is frequently used to designate the process of extracting useful information from large databases. In this chapter, we adopt a slightly different view, which is identical to the one expressed . In this view, the term knowledge discovery in databases (KDD) is used to denote the process of extracting useful knowledge from large data sets. Data mining, by contrast, refers to one particular step in this process. Spherically, the data mining step applies so-called data mining techniques to extract patterns from the data. Additionally, it is preceded and followed by other KDD steps, which ensure that the extracted patterns actually correspond to useful knowledge. Indeed, without these additional KDD steps, there is a high risk of finding meaningless or uninteresting patterns . In other words, the KDD process uses data mining techniques along with any required pre- and post-processing to extract high-level knowledge from low-level data. In practice, the KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user. Here, we broadly outline some of the most basic KDD steps: 1. Understanding the application domain: First is developing an understanding of the application domain, the relevant background knowledge, and the specific goals of the KDD endeavor. 2. Data integration and selection: Second is the integration of multiple data and the selection of the subset of data that is relevant to the analysis task. 3. Data mining: Third is the application of specific algorithms for extracting patterns from data. 4. Pattern evaluation: Fourth is the interpretation and validation of the discovered patterns. The goal of this step is to guarantee that actual knowledge is being discovered. 5. Knowledge representation: This step involves documenting and using the discovered knowledge. In other words, data mining emphasizes the efficient discovery of simple, but understandable models that can be interpreted as interesting or useful knowledge. In fact, data mining is just a step in the KDD process. As such, it has to contribute to the overall goal of knowledge discovery. Some Data Mining Techniques: Data mining techniques essentially are pattern discovery algorithms. Some techniques such as association rules are unique to data mining, but most are drawn from related fields such as machine learning or pattern recognition. In this section, we introduce four well-known data mining techniques that have been widely used in intrusion detection. A broader and more detailed treatment of data mining techniques can be found elsewhere. A potential source of confusion is that different data mining techniques assume different input data representations. For example, association rules have historically been discussed under the assumption that the input data is represented as a set of transactions.Later, association rule mining over relational databases has been investigated. Depending on the input data representations (sets of transactions versus relational databases), the association rule concept is presented differently. A related problem is that there are many different ways to represent the same data set in a relational database. In practice, the available input data does not necessarily follow this format. Then, it is the responsibility of the second KDD step to transform the available data into the format required by the data mining techniques.

- Association Rules
- Frequent Episode Rules
- Classification
- Clustering

5. CLASSIFICATION OF TECHNIQUES

In a classification task in machine learning, the task is to take each instance of a dataset and assign it to a particular class. A classification based IDS attempts to classify all traffic as either normal or malicious. The challenge in this is to minimize the number of false positives (classification of normal traffic as malicious) and false negatives (classification of malicious traffic as normal). Five general categories of techniques have

been tried to perform classification for intrusion detection purposes:

- INDUCTIVE RULE GENERATION
- GENETIC ALGORITHMS
- NEURAL NETWORKS
- PASSIVE SYSTEM V/S REACTIVE SYSTEM

6. LITERATURE SURVEY

Chandrakant Namdev et. al [2015] proposed that data mining is the process of extracting valid,previously known & comprehensive datasets for the future decision making. In this paper we present the mechanism to improve the efficiency of the IDS using streaming data mining technique. We apply four stream data classification algorithms on NSL-KDD datasets and compare their results based on the comparative analysis of their results best method is found out for efficiency improvent of IDS.

Aditya Shrivastava1 et.al [2013] we proposed a hybrid model for feature selection and intrusion detection. Feature selection is important issue in intrusion detection. The selection of feature in attack attribute and normal traffic attribute is challenging task. The selection of known and unknown attack is also faced a problem of classification. PCNN is dynamic network used for the process of feature selection in classification. The dynamic nature of PCNN select attribute on selection of entropy. The attribute entropy is high the feature value of PCNN network is selected and the attribute value is low the PCNN feature selector reduces the value of feature selection. After selection of feature the Gaussian kernel of support vector machine is integrated for classification. Our detection rate is very high in compression of other neural network model such as RBF neural network and SOM network.

JAYSHRI R. PATEL et. al [2013] proposed a Decision Trees are considered to be one of the most popular approaches for representing classifier for various disciplines such as statistics, machine learning and data mining. Classification of Intrusion detection, according to their features into either intrusive or non intrusive class is a widely studied problem. Decision trees are useful to detect intrusion from connection records. In this paper, we evaluate the performance of various decision tree classifiers for classifying intrusion detection data. The aim of this paper is to investigate the performance of various decision tree classifiers for ranked intrusion detection data.

Venkata Suneetha Takkellapati et. al [2012] proposed As the cost of the data processing and Internet accessibility increases, more and more organizations are becoming vulnerable to a wide range of cyber threats. Most current offline intrusion detection systems are focused on unsupervised and supervised machine learning approaches. Existing model has high error rate during the attack classification using support vector machine learning algorithm. Besides, with the study of existing work, feature selection techniques are also essential to improve high efficiency and effectiveness. Performance of different types of attacks detection should also be improved and evaluated using the proposed approach.

Saurabh Mukherjee, Dr., Neelam Sharma et. al[2012] proposed Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems. Today most of the intrusion detection approaches focused on the issues of feature selection or reduction, since some of the features are irrelevant and redundant which results lengthy detection process and degrades the performance of an intrusion detection system (IDS). The purpose of this study is to identify important reduced input features in building IDS that is computationally efficient and effective. For this we investigate the performance of three standard feature selection methods using Correlation-based Feature Selection, Information Gain and Gain Ratio. In this paper we propose method Feature Vitality Based Reduction Method, to identify important reduced input features. We apply one of the efficient classifier naive bayes on reduced datasets for intrusion detection.

Carlos A. Cataniaa et .al [2012] Automatic network intrusion detection has been an important research topic for the last 20 years. In that time, approaches based on signatures describing intrusive behavior have become the de-facto industry standard. Alternatively, other novel techniques have been used for improving automation of the intrusion detection process. In this regard, statistical methods, machine learning and data mining techniques have been proposed arguing higher automation capabilities than signature-based approaches. However, the majority of these novel techniques have never been deployed on real-life scenarios. The fact is that signature-based still is the most widely used strategy for automatic intrusion detection. In the present article we survey the most relevant works in the field of automatic network intrusion detection. In contrast to previous surveys, our analysis considers several features required for truly deploying each one of the reviewed approaches.

Ahmed Patel et .al [2013] The distributed and open structure of cloud computing and services becomes an attractive target for potential cyber-attacks by intruders. The traditional Intrusion Detection and Prevention Systems (IDPS) are largely inefficient to be deployed in cloud computing environments due to their openness and specific essence. This paper surveys, explores and informs researchers about the latest developed IDPSs and alarm management techniques by providing a comprehensive taxonomy and investigating possible solutions to detect and prevent intrusions in cloud computing systems.

Chirag Modia et .al [2013] In this paper, we survey different intrusions affecting availability, confidentiality and integrity of Cloud resources and services. Proposals incorporating Intrusion Detection Systems (IDS) and Intrusion Prevention

Systems (IPS) in Cloud are examined. We recommend IDS/IPS positioning in Cloud environment to achieve desired security in the next generation networks.

7. PROBLEM FORMULATION

In the research work intrusion detection different problems are faced. A hybrid model for feature selection and intrusion detection is important issue in intrusion detection. The selection of feature in attack attribute and normal traffic attribute is challenging task. There are different problems that are given below:

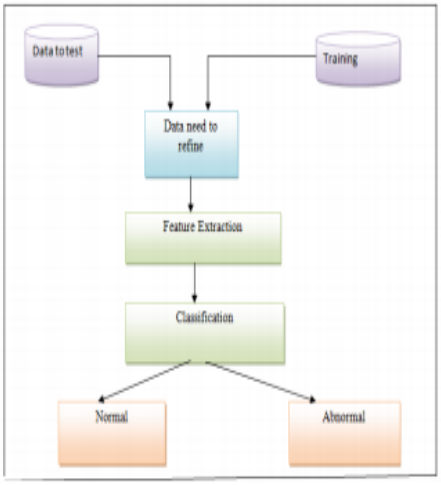
- There are selection of known and unknown attack is also faced a problem of classification. There is multiclass problem during the classification of data.
- Intrusion detection is a problem of transportation infrastructure protection owing to the fact that computer networks are at the core of the operational control of much of the nation's transportation.
- The feature ranking and selection problem for intrusion detection is similar in nature to various engineering

Another major problem that is also faced in the intrusion detection that are given below:

1. Security
2. Authentication
3. Attackers

8. METHODOLOGY

This is to detect the intrusion from network. It is based upon weka tool. Their are the programmable files containing the information about the dataset. The Intrusion detection system deals with large amount of data which contains various irrelevant and redundant features resulting in increased processing time and low detection rate.



Therefore feature selection plays an important role in intrusion detection. There are various feature selection methods proposed in literature by different authors. In this a comparative analysis of different feature selection methods are presented on KDDCUP'99 benchmark dataset and their performance are evaluated in terms of detection rate and TP and FP and Precision and Recall.

Theoretical steps for the proposed methodology:

- Gather the audit data i.e., data to test from Kdd'99.
- Similarly the gather the infonnation for the training data set.
- Use k- means to make dataset to be redundant and form k- clusters. This step is basically to encourage the redundancy.
- Input for the SVM algorithm is ready.
- Calculate weight, of each attribute.
- Now using association rule clusters will be classified.
- Those classified clusters are either anomaly or normal.

9. CONCLUSION

Today most of the intrusion detection approaches targeted on the problems of feature choice or reduction, since a number of the options area unit impertinent associate degreed redundant which ends long detection method and degrades the performance of an intrusion detection system (IDS). the aim of this study is to spot necessary reduced input options in building IDS that's computationally economical and effective. For this we tend to investigate the performance of 3 commonplace feature choice ways exploitation Correlation-based Feature choice, info Gain and Gain magnitude relation. during this paper intrusion detection is reviewed with the assistance completely different of various} analysis papers that was antecedently done by different researchers. during this paper KDDCup's ninety nine dataset is employed to implement this add the long run.

REFERENCES

[1] C. C. Aggarwal, P. Yu, Outlier Detection for High Dimensional Data, Proceedings of the ACM SIGMOD Conference, 2001.
[2] D. Barbara, N. Wu, S. Jajodia, Detecting Novel NetworkIntrusions Using Bayes Estimators, First SIAM Conferenceon Data Mining, Chicago, IL, 2001
[3] D.E. Denning, An Intrusion Detection Model, IEEE Transactionson Software Engineering, SE-13:222- 232, 1987.
[4] E. Bloedorn, et al., Data Mining for Network IntrusionDetection: How to Get Started, MITRE Technical Report,August 2001
[5] E. Knorr, R. Ng, Algorithms for Mining Distance-based Outliers in Large Data Sets, Proceedings of the VLDB Conference, 1998.
[6] F. Provost and T. Fawcett, Robust Classification for ImpreciseEnvironments, Machine Learning, vol. 42/3, pp. 203-231, 2001
[7] H.S. Javitz, and A. Valdes, The NIDES Statistical Component:Description and Justification, Technical Report,Computer Science Laboratory, SRI International, 1993
[8] J. Luo, Integrating Fuzzy Logic With Data Mining Methodsfor Intrusion Detection, Master's thesis, Department ofComputer Science, Mississippi State University, 1999.

[9] Lazarevic, N. Chawla, L. Hall, K. Bowyer, SMOTEBoost:Improving the Prediction of Minority Class inBoosting, AHPCRC Technical Report, 2002.
[10] M. Joshi, V. Kumar, R. Agarwal, Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements, First IEEE International Conference on Data Mining, San Jose, CA, 2001.
[11] M.Joshi, R. Agarwal, V. Kumar, Predicting Rare Classes:Can Boosting Make Any Weak Learner Strong?, Proceedingsof Eight ACM Conference ACM SIGKDD InternationalConference on Knowledge Discovery and Data Mining,Edmonton, Canada, 2002.
[12] M. Joshi, R. Agarwal, V. Kumar, PNrule, Mining Needlesin a Haystack: Classifying Rare Classes via Two-PhaseRule Induction, Proceedings of ACM SIGMOD Conferenceon Management of Data, May 2001.
[13] M. Joshi, V. Kumar, CREDOS: Classification using RippleDown Structure (A Case for Rare Classes), in review.
[14] M. M. Breunig, H.-P. Kriegel, R. T. Ng, J. Sander, LOF:Identifying Density-Based Local Outliers , Proceedings ofthe ACM SIGMOD Conference, 2000.
[15] P.C. Mahalanobis, On Tests and Meassures of Groups Divergence,International Journal of the Asiatic Society ofBenagal, 26:541, 1930.
[16] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. P.Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod,R. K. Cunningham, and M. A. Zissman, EvaluatingIntrusion Detection Systems: The 1998
[17] DARPA OfflineIntrusion Detection Evaluation, Proceedings DARPAInformation Survivability Conference and Exposition(DISCEX) 2000, Vol 2, pp. 12-26, IEEE Computer SocietyPress, Los Alamitos, CA, 2000
[18] R. P. Lippmann, R. K. Cunningham, D. J. Fried, I. Graf, K.R. Kendall, S. W. Webster, M. Zissman, Results of the1999 DARPA Off-Line Intrusion Detection Evaluation,Proceedings of the Second International Workshop on RecentAdvances in Intrusion Detection (RAID99), West Lafayette,IN, 1999.
[19] Successful Real-Time Security Monitoring, Riptech Inc.white paper, September 2001.
[20] S. Manganaris, M. Christensen, D. Serkle, and K. Hermix,A Data Mining Analysis of RTID Alarms, Proceedings ofthe 2nd International Workshop on Recent Advances in IntrusionDetection (RAID 99), West Lafayette, IN, September1999
[21] S. Ramaswamy, R. Rastogi, K. Shim, Efficient Algorithms for Mining Outliers from Large Data Sets, Proceedings of the ACM SIGMOD Conference, 2000.
[22] V. Barnett, T. Lewis, Outliers in Statistical Data, JohnWiley and Sons, NY 1994
[23] W. Lee, S. J. Stolfo, Data Mining Approaches for IntrusionDetection, Proceedings of the 1998 USENIX Security Symposium,1998.