International Journal of Advanced Computer Research (ISSN (print):2249-7277 ISSN (online):2277-7970)
Volume-3 Number-2 Issue-10 June-2013

Design of Intrusion Detection Model Based on FP-Growth and Dynamic Rule Generation with Clustering

Manish Somani[1], Roshni Dubey[2]

Abstract

Intrusion Detection is the process used to identify intrusions. If we think of the current scenario then several new intrusion that cannot be prevented by the previous algorithm, IDS is introduced to detect possible violations of a security policy by monitoring system activities and response in all times for betterment. If we detect the attack type in a particular communication environment, a response can be initiated to prevent or minimize the damage to the system. So it is a crucial concern. In our framework we present an efficient framework for intrusion detection which is based on Association Rule Mining (ARM) and K-Means Clustering. K-Means clustering is use for separation of similar elements and after that association rule mining is used for better detection. Detection Rate (DR), False Positive Rate (FPR) and False Negative Rate (FNR) are used to measure performance and analysis experimental results.

Keywords

ARM, K-Means, DR, FPR, FNR

1. Introduction

Nowadays, as information systems are more open to the Internet, the importance of secure networks is extremely increased. New clever Intrusion Detection Systems (IDSs) which are based on sophisticated algorithms rather than current signature-base detections are in demand. There is frequently the need to update an installed Intrusion Detection System (IDS) due to new attack methods or upgraded computing environments. Since many present Intrusion Detection Systems are constructed by manual encoding of expert knowledge, changes to them are costly and slow. In data mining based intrusion detection system, we should make employ of particular domain knowledge in relation to intrusion detection in order to efficiently extract relative rules from large amounts of records. This paper proposes new ensemble boosted decision tree approach for intrusion detection system. Experimental outcome shows better results for detecting intrusions as compared to others existing methods.

An Intrusion Detection System (IDS) is expected to identify cruel behaviors that menace the integrity, secrecy and availability of network resources. It can help to detect suspicious activities in the network. Likewise the major conventional detection techniques: misuse detection and anomaly detection. Misuse detection systems [1],[2] are most widely used and they detect intruders with known patterns and signatures used to identify attacks that consist of various fields of network based packet information, similar to source address, destination address, and source and destination ports. This model needs continuous updating because of Adhoc network, but they have a virtue of having very low false positive rate. Anomaly detection systems identify deviations from normal user behavior and alert to potential unknown or novel attacks without having any prior knowledge of them. They exhibit higher rate of false alarms and hence the need for a mechanism to reduce false approach arises. The traditionally detection methods are not suitable in fast growing network environment, and it also not considered for focusing massive knowledge of engineering and information security task, experts are involved to analyzes the various mathematical methods for anomaly detection and also concentrate on various sequence code and patterns to detect the misuse detection, everything is computerized by machine learning code to detect the detection. These methods are not suitable for current dynamic nature environment. Basic ideas for [3], [4] are enabled in to system to record the activities like System events, legitimate activities and intrusion behavior. But this type of measures need advanced tools and techniques to detect suspicious activities. Both detection methods have been extensively studied by the research community for many years.

KDD99Cup and DARPA98 datasets [5],[6] provided by MIT Lincoln Laboratories are widely used as training and testing datasets for the evaluation of IDSs KDDCUP'99 intrusion detection dataset contains a standard dataset andtraining data to be audited .It analyzes the connections of 41features. All the connections can be divided into 5 categories including normal network connection, and other four

categories are Denial of Service Attack (DOS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing Attack.

The remaining of this paper is organized as follows. In Section 2 we discuss about data mining and intrusion detection. Literature Surveyin section 3. In section 4 we discuss about the proposed framework. The conclusions are given in Section 5. Lastly references are given.

2. Data Mining and Intrusion Detection

The successful data mining techniques are themselves not enough to create deployable IDSs. . In spite of the promise of better detection performance and generalization ability of data mining-based IDSs, there are a few inherent difficulties in the implementation and deployment of these systems. In this paper, we discuss a number of problems inherent in developing and deploying a real-time data mining-based IDS and present an overview of our research, which talk about these problems. These problems are independent of the real learning algorithms or models used by IDS and must be overcome in order to implement data mining methods in a deployable system [7] and [8] and [9].

Association Rule

Association rules mining identifies associations (patterns or relations) among database attributes and their values. It is a pattern discovery method which does not serve to solve classification problems (it does not classify samples into some target classes) nor prediction problems (it does not predict the development of the attribute values). Association rules mining generally searches for any associations among any attributes present in the database. Association rule (AR) is commonly understood as an implication X $\rightarrow$Y in a transaction database D = {t1…….. tm}. Each transaction ti$\in$D contains a subset of items I = {i1.....in}. X and Y are disjoint item sets, it holds X; Y $\subseteq$I and X $\cap$ Y = Φ. The left hand side of this implication is called precursor, the right hand side is referred to as consequent. The transaction database D can as well be viewed as a boolean dataset where the boolean values of attributes in records express occurrence of items in transactions.

Association rule mining problem poses the question of efficiency. The number of potential rules X $\rightarrow$ Y defined by X $\subseteq$ Ix $\in$ {ix1;…… ;ixn}, Y $\subseteq$Iy$\in$ {iy1;…. ; iym}, where Ix and Iy are disjoint, is equal to 2(m+n). When general datasets are considered, the AR mining problem is known to be NP-complete. In restricted cases, for illustration in sparse boolean datasets (where it holds all ti$\in$ D; |ti| <= O(log|I|)) lower complexity bounds have been proved to clutch. Finding rules in quantitative data further strengthens importance of efficiency. A raise in the number of values that can be associated with any given variable increases the number of rules exponentially, thus causing execution time to raise significantly [10] and [11].

In typical applications of data mining to intrusion detection, detection models are manufacture off line because the learning algorithms must process tremendous amounts of archived audit data. These models can naturally be utilized for off line intrusion detection (i.e., analyzing audit data offline after intrusions have run their course). Effectual intrusion detection should happen in real-time, as intrusions take place, to reduce security compromises. Now we discuss the approaches to make data mining-based ID models work efficiently for real-time intrusion detection. In contrast to offline IDSs, a key objective of real-time IDS is to detect intrusions as near the beginningas possible. Therefore, the efficiency of the detection model is a very important consideration. For the reason that the data mining-based models are computed using off line data, they implicitly suppose that when an event is being inspected (i.e., classified using an ID model), each activities related to the event have completed so that all features have meaningful values available for model checking. Unfortunately, DoS attacks, which typically generate a large amount of traffic in a very short period time, are frequently used by intruders to first overload an IDS, and employ the detection delay as a window of opportunity to quickly perform their malicious intent. For illustration, they can even seize control of the host on which the IDS lives, so eliminating the effectiveness of intrusion detection altogether. It is essential to examine the time delay associated with computing each feature in order to speed up model evaluation. From the perspective of cost analysis, the effectiveness of an intrusion detection model is its calculation cost, which is the sum of the time delay of the features used in the model [12] and [9].

3. Literature Survey

In 2010,G. Schaffrath et al. [13] provide a survey of current research in the area of flow-based intrusion detection. The survey begins with a motivation why

flow-based intrusion detection is needed. The concept of flows is give details, and relevant standards are identified. The paper gives a classification of attacks and defense techniques and shows how flow-based techniques can be used to detect scans, worms, Botnets and (DoS) attacks.

In 2012, R.Venkatesan et al. [14] survey and analysis that data mining techniques have been successfully applied in many fields like Network Management, Education, Science, Business, industrialized, Process control, and Fraud Detection. Data Mining for IDS is the method which can be used mainly to identify unknown attacks and to raise alarms when security violations are detected. In 2012, Sneha Kumari et al. [15] suggest that the Over the past several years, the Internet atmosphere has become more complex and untrusted. Enterprise networked systems are unavoidably exposed to the increasing threats posed by hackers as well as malicious users internal to a network. IDS technology is one of the vital tools used now-a-days, to counter such threats. Authors also provide the comparison study which is based on artificial neural network (ANN),Bayesian network classifier (BNC), Support Vector Machine (SVM) and Decision Tree (DT).

In 2012, Vineet Richariya et al. [16] analyzed the performance and applicability of the well knows IDS system based on mobile agent with their advantages and disadvantages. Mobile agent is efficient way to find out the intruder in distributed system. The main features of mobile agents are intelligence and mobility which is the core motivation to us to designed cost. The aim of their review work is to help to select appropriate IDS systems as per their requirement and application.

In 2012, Deepak Rathore et al. [17] proposed an ensemble Cluster Classification technique using SOM network for detection of mixed variable data generated by malicious software for attack purpose in host system. In their methodology SOM network control the iteration of distance of different parameters of ensemble their experimental result which show better empirical evaluation on KDD data set 99 in comparison of existing ensemble classifier. In 2012, LI Yin–huan [18] focuses on an improved FP-Growth algorithm. According to author Preprocessing of data mining can increase efficiency on searching the common prefix of node and reduce the time complexity of building FP-tree. Foundation on the improved FP Growth algorithm and other data mining methods, an intrusion detection model is carried out by authors. Their experimental results are effective and feasible.

In 2012, P. Prasenna et al. [19] suggested that in conventional network security simply relies on mathematical algorithms and low counter measures to taken to prevent intrusion detection system, although nearly all of this approaches in terms of theoretically challenged to implement. Authors suggest that instead of generating large number of rules the evolution optimization techniques like Genetic Network Programming (GNP) can be used .The GNP is based on directed graph. They focus on the security issues related to deploy a data mining-based IDS in a real time environment. They generalize the problem of GNP with association rule mining and propose a fuzzy weighted association rule mining with GNP framework suitable for both continuous and discrete attributes.

4. Proposed Work

In our proposed work we first access the data from the valid database like KDD CUP 99 database is accessed. Then the data is pre-processed. Pre-process phase is like the data audit phase. Because the data we taken are not necessary support the properties our framework. So in pre-processing phase we first make it compatible to the framework we used. Then we also check the redundant data so that we only process the meaningful data, means the set contains unique items, so the processing overhead is reduced. Then the data is applied for clustering as shown in figure 1.

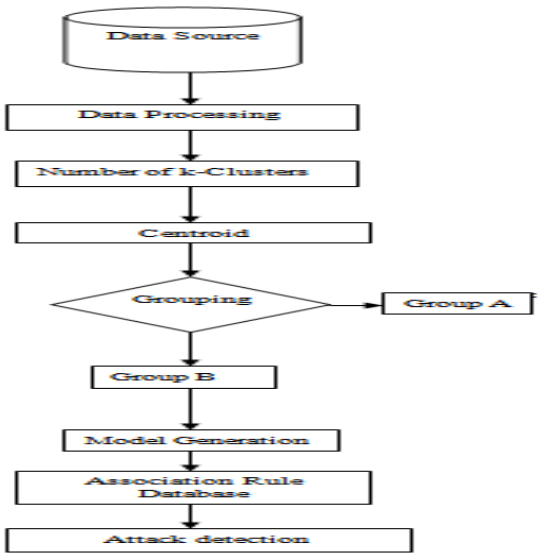

Figure 1: Flowchart of working Process

Then we apply K-Means clustering for separation for the alike elements.

The categorization of objects into various groups or the partitioning of data set into subsets so that the data in each of the subset share a common feature, frequently the proximity with regard to some defined distance measure, is known as Clustering. The clustering problem has been addressed in numerous contexts besides being proven beneficial in many applications. Clustering medical data into little yet meaningful clusters can aid in the discovery of patterns by supporting the extraction of numerous appropriate features from each of the clusters thereby introducing structure into the data and aiding the application of conventional data mining techniques. Numerous methods are available in the literature for clustering. We have utilized the renowned K-Means clustering algorithm in our approach.

The k-means algorithm is one of the extensively recognized clustering tools that are applied in a variety of scientific and industrial applications. K-means groups the data in accordance with their characteristic values into K distinct clusters. Data classify into the same cluster have identical feature values. K, the positive integer indicate the number of clusters, wants to be provided in advance.

The steps occur in a K-means algorithm are given subsequently:

1. K points denoting the data to be clustered are placed into the space. These points indicate the primary group centroids.
2. The data are assigned to the group that is adjacent to the centroid.
3. The positions of all the K centroids are recalculated as soon as all the data are assigned.

Steps 2 and 3 are reiterated until the centroids stop moving any further. This outcomes in the Segregation of data into groups from which the metric to be minimized can be deliberated. The preprocessed software estimation data warehouse is clustered using the K-means algorithm with K value as 4. Because we need the separation based on four different object oriented parameters that is class, object, inheritance and dynamic behavior.

Then we apply association rule mining by non-candidate rule generation.

5. Conclusion

Traditional Data Mining techniques operate on structured data such as corporate databases; this has been an active area of research for many years. Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems. Intrusion detection is an area growing in relevance as more and more sensitive data are stored and processed in networked systems. An intrusion detection system (IDS) monitors networked devices and looks for anomalous or malicious behavior in the patterns of activity in the audit stream. The application of data mining techniques to IDS is one important direction for future development of intrusion detection. Selecting appropriate data mining algorithms and designing IDS model are effective measures in order to improve system detection performance. In this paper we discuss an efficient framework for intrusion detection based on data mining.

References

[1] J. G.-P. A. El Semaray, J. Edmonds, and M.Papa, “Applying datamining of fuzzy association rules to network intrusion detection,” presented at the IEEE Workshop Inf., United States Military Academy,West Point, NY, 2006.

[2] A. S. S. Forrest, S. A. Hofmeyr, and T. A. Longstaff, “A sense of self for unix processes,” presented at the IEEE Symp. Secur.Privacy, Los Alamitos, CA, 1996.

[3] J. Luo, “Integrating fuzzy logic with data mining methods for intrusion detection,” Master’s thesis, Dept. Comput.Sci., Mississippi State Univ., Starkville, MS, 1999.

[4] Lippmann, Richard P., David J. Fried, Isaac Graf, Joshua W. Haines, Kristopher R. Kendall, David McClung, Dan Weber et al. "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation." In DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings, vol. 2, pp. 12-26. IEEE, 2000.

[5] KDDCUP 1999 data [Online]. Available: kdd.ics.uci.edu/databases/ kddcu p99/kddcup99.html.

[6] Darpa Intrusion Detection datasets [Online]. Available:www.ll.mit.edu/missioncommunicatio ns/ist/corpora/ideval/data/index.html.

[7] RakeshShrestha, Kyong-Heon Han, Dong-You Choi, Seung-Jo Han, “A Novel Cross Layer Intrusion Detection System in MANET”, 2010 24th IEEE International Conference on Advanced

Information Networking and Applications, pp 647-656.

[8] Shaik Akbar, Dr.K.NageswaraRao and Dr.J.A.Chandulal, "Intrusion Detection System Methodologies Based on Data Analysis", International Journal of Computer Applications (0975 – 8887) Volume 5– No.2, August 2010, pp 10-20.

[9] Abhinav Srivastava, Shamik Sural and A.K. Majumdar, "Database Intrusion Detection using Weighted Sequence Mining", Journal Of Computers, Vol. 1, No. 4, July 2006, pp 8-17.

[10] PrakashRanganathan, Juan Li, Kendall Nygard, "A Multiagent System using Associate Rule Mining (ARM), a collaborative filtering approach", IEEE 2010, pp- v7 574- 578.

[11] Prof Thivakaran.T.K, Rajesh.N, Yamuna.P, PremKumar.G, "Probable Sequence Determination Using Incremental Association Rule Mining And Transaction Clustering", IEEE 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, pp 37-41.

[12] Preetee K. Karmore , Smita M. Nirkhi, "Detecting Intrusion on AODV based Mobile Ad Hoc Networks by k-means Clustering method of Data Mining", International Journal of Computer Science and Information Technologies, Vol. 2 (4) , 2011, pp 1774-1779.

[13] G.Schaffrath,R. Sadre,C. Morariu,A.Pras and B.Stiller, "An Overview of IP Flow-Based Intrusion Detection", Communications Surveys & Tutorials, IEEE 2010.

[14] R.Venkatesan, R. Ganesan and A. Arul Lawrence Selvakumar, "A Comprehensive Study in Data Mining Frameworks for Intrusion Detection", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-4, Issue-7, December-2012.

[15] SnehaKumari, ManeeshShrivastava, "A Study Paper on IDS Attack Classification Using Various Data Mining Techniques", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5, September-2012.

[16] Vineet Richariya, Uday Pratap Singh, Renu Mishra,"Distributed Approach of Intrusion Detection System: Survey", International Journal of Advanced Computer Research (ISSN (IJACR), Volume-2, Number-4, Issue-6 December-2012.

[17] Deepak Rathore, Anurag Jain, "Design Hybrid method for intrusion detection using Ensemble cluster classification and SOM network", International Journal of Advanced Computer Research (IJACR), Volume-2, Number-3, Issue-5, September-2012.

[18] LI Yin–huan , "Design of Intrusion Detection Model Based on Data Mining Technology", International Conference on Industrial Control and Electronics Engineering, 2012.

[19] P. Prasenna,R. Krishna Kumar, A.V.T Raghav Ramana and A. Devanbu "Network Programming And Mining Classifier For Intrusion Detection Using Probability Classification",Pattern Recognition, Informatics and Medical Engineering, March 21-23, 2012.