IJERMT
International Journal of Engineering Research & Management Technology ISSN: 2348-4039
May-2015 Volume 2, Issue-3
Email: editor@ijermt.org www.ijermt.org

DATA MINING IN NETWORK SECURITY TECHNIQUES & TOOLS

(A Survey and Comparative Analysis of Data Mining Techniques for Network Security)

PRADEEP KUMAR
M.Tech Student
Deptt .Computer Science & Engineering
Monad University
Hapur U.P. India

MAHESH KUMAR
Assistant Professor
Deptt. Computer Science & Engineering
Monad University
Hapur U.P. India

Abstract:

In this digital age we can't imagine the world without communication. The human beings need to exchange information for various purposes, securing the communication is a vast challenge due to the rising threats and attacks against network security. Despite of growing information technologies widely security has remained one challenging area for computers & Networks. In information security, intrusion detection is the act of detecting action that attempt to compromise the confidentiality, integrity, availability of resources. Currently many researchers have focused on intrusion detection. Data mining is one of the technologies applied to the intrusion detection to invent a new pattern from a massive network data as well as to reduce the strain of the manual compilations of the instruction and normal behaviour pattern. This article reviews the current state of art data mining techniques, comprise various data mining techniques used to implement an intrusion detection system such as "decision trees", "artificial neural network", "naïve Bayes", "support vector machine" ,"k-nearest neighbour algorithms" by highlighting advantages and disadvantages of each of the techniques. And finally a discussion of the future technologies and methodologies which promises to enhance the ability of computer system to detect intrusion is provided and current research challenges are pointed out in the field of intrusion detection system. This will be beneficial to the academician, industrialists, and students who incline towards research and development in the area of data mining in network security.

Keywords: Learning, Intrusion Detection, Unsupervised Learner, Data Mining, Network Security.

Introduction

Intrusion detection technique is technology designed to observe computer activities for the purpose of finding security violations. The security of a computer system is compromised when an intrusion takes place. Intrusion detection is the process of identifying and responding to malicious activity targeted at computing networking resources [1].

The threats are classified based on their behaviour such as:

Leakage:

Unauthorized access of information available in the network [2]

Tampering:

Modifying the information without permission of the author

Vandalism:

Making mal function over a normal execution of a system. The various types of attacks such as: **eaves dropping**:

Collecting the replica information without obtaining permission to the arbiter. Masquerading: Making conversation using through other identity without permission of others.

Man-In-The-Middle Attack:

Is a one type of messaging interfering in which an attacker interrupt the very first message in an exchange of encrypted keys to establish a secure channel [3][4].

Replying:

This is one type of attack that stores interrupt message then sends these messages later. This attack may be effective even with authenticated and encrypted messages [3].Denial of services makes the transmission channels and systems as busy as possible by sending garbage data for denying the service [5].The knowledge about these attacks is acquired from the huge volume of network data with data mining tools. This knowledge facilitates the security system to identify the attackers or hackers based on their behaviour in a network. The behaviour of the attackers and hackers are studied and identified by two types of learning strategies namely unsupervised and supervised learning. As network based computer system play increasingly vital roles in modern society. They have become intrusion detection a system provides following there essential security functions:

Data Confidentiality:

Information that is being transferred through the network should be accessible only to those that have been properly authorized.

Data Integrity:

Information should maintain then integrity from the moment they are transmitted to the moment they are actually received no corruption or data loss is accepted either from the random events or malicious activity.

Data Availability:

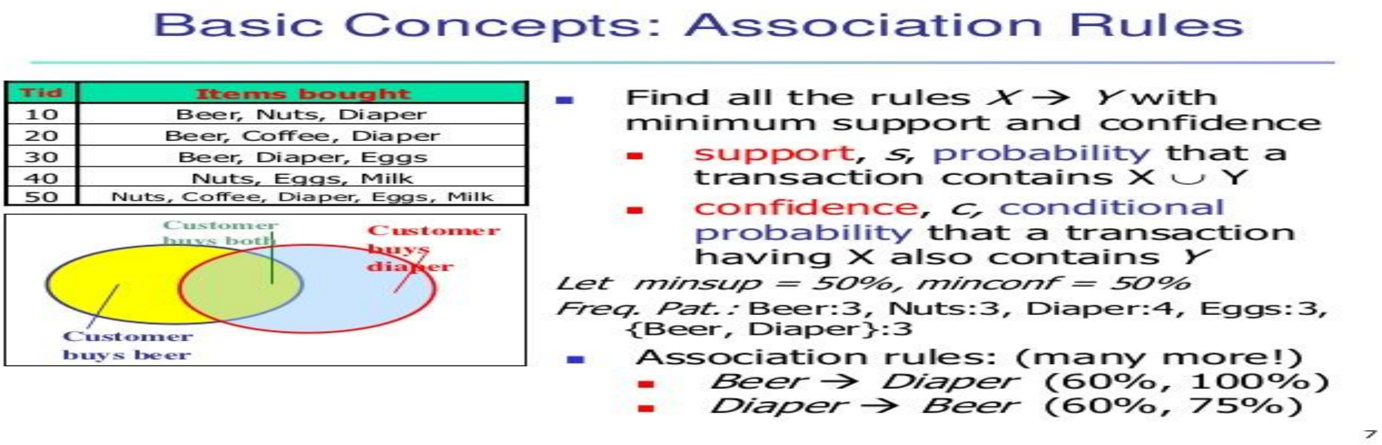
The network or system resources that ensure that it is accessible and usable upon demand by an authorized system user . This paper proposes an unsupervised learning based intrusion detection system that utilizes the advantages of unsupervised learning prediction techniques. Also a detailed discussion on the step by step procedure of building a predictive model using data mining tools is presented

Literature Survey

The successful data mining techniques are themselves not enough to create deployable IDSs. Despite the promise of better detection performance and generalization ability of data mining-based IDSS, there are some inherent difficulties in the implementation and deployment of these systems. In this paper, we discuss several problems inherent in developing and deploying a real-time data mining-based IDS and present an over view of our research, which addresses these problems. These problems are independent of the actual learning algorithms or models used by IDS and must be overcome in order to implement data mining methods in deployable systems.[17][18][19]

Association Rule

Association rules mining identifies association (patterns or relations) among database attributes and their values. it is a pattern discovery technique 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 researches for any associations among any attributes present in the database. As follow:



Basic Concepts: Association Rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - support, s, probability that a transaction contains $X \cup Y$
  - confidence, c, conditional probability that a transaction having X also contains Y

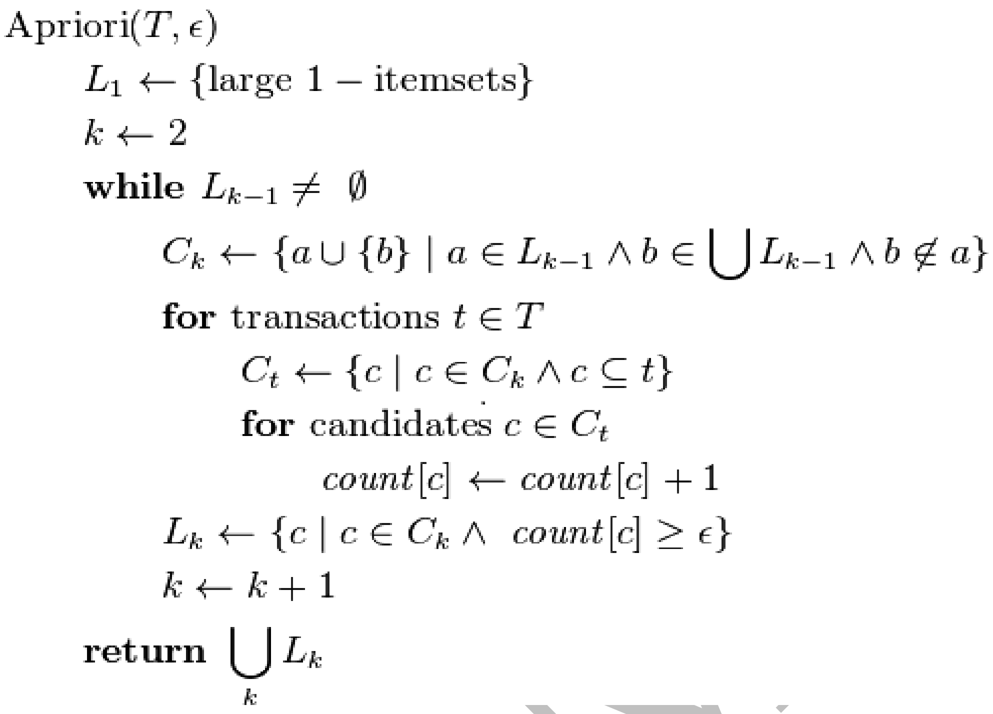
Let minsup = 50%, minconf = 50%

Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

- Association rules: (many more!)
  - *Beer* → *Diaper* (60%, 100%)
  - *Diaper* → *Beer* (60%, 75%)

7

Association rule algorithm: [24]



```
Apriori($T, \epsilon$)
    $L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\}$
    $k \leftarrow 2$
    while $L_{k-1} \neq \emptyset$
        $C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\}$
        for transactions $t \in T$
            $C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$
            for candidates $c \in C_t$
                $count[c] \leftarrow count[c] + 1$
        $L_k \leftarrow \{c \mid c \in C_k \wedge count[c] \geq \epsilon\}$
        $k \leftarrow k + 1$
    return $\bigcup_k L_k$
```

Classification Techniques:

Classification is the process of learning a function that maps data objects to subset of a given class set. Therefore, a classifier is trained with a labelled set of training objects, specifying each class. There are two goals of classifications: finding a good general mapping that can predict the class of so far unknown data object with high accuracy. For this goal, the classification is a mere function. to achieve this goal, the classifier has to decide which of the characteristics of the given training instances are typical for the complete class and which characteristics are specific for single object in the training set. the other goal of classification is to find a compact and understandable class model for each of the classes. A class model should give an explanation why the given objects belong to a certain class and what is typical for the member of a given class. The class model should be as compact as possible because the more compact a model is, the more general it is. Furthermore, small and simple class models are easier to understand and contain less distracting information. Of course, a good classifier should serve both purposes, but for most practical applications finding an accurate mapping is more important than developing understandable class model. Thus, multiple techniques are used to classify objects that do not offer an understandable class model.

Clustering Techniques:

Clustering analysis is a very broad field and the number of available methods and their variations can be overwhelming. A good introduction to numerical clustering can be found in cluster analysis or in cluster classifications. A more up to date view of clustering in the context of data mining is available in data mining concepts and techniques. As:

Partitioning Methods:

Partitioning clustering method divide the input into disjoint subsets attempting to find a configuration which maximizes some optimality criterion. Because enumeration of all possible subsets of the input is usually computationally infeasible, partitioning clustering employs an iterative improvement procedure which moves objects between clusters until the optimality criterion can no longer be improved. This most popular partitioning algorithm is the K-Means, we define a global objective function and iteratively move objects between partitions to optimize this function. The objective function is usually a sum of distances (or sum of squared distance) between objects and their cluster's centres and the objective is to minimize it the representation of a cluster can be an average of its elements or a mean point. In the latter case we call the algorithm K-Medoids. Given the number of clusters K a priori, a generic K-Means procedure is implemented in four steps:

1. Partition object into K nonempty subset.

2. Compute representation of canters for current clusters.
3. Assign each object to the closet cluster.
4. Repeat from step 2 until no more reassignments occur.

Hierarchical Methods:

A family of hierarchical clustering methods can be divided into agglomerative and divisive variants. Agglomerative hierarchical clustering initially places each object in its own cluster and then iteratively combines the closet cluster merging their content. The clustering process is interrupted at some point. Leaving a dendrogram with a hierarchy a cluster. Many variant hierarchical methods exist, depending on the procedure of locating pairs of cluster to be merged. in the single link method, the distance between clusters is the minimum distance between any pair of elements drawn from these clusters. in the complete link it is the maximum distance and in the average link. It is correspondingly an average distance (a discussion of other merging methods can be found. Each of these has a different computational complexity and runtime behaviour. Single link method is known to follow “Bridges” of noise and link elements in distant clusters). Complete link method is computationally more demanding. But is known to produce more sensible hierarchies. Average link method is a trade-off between speed and quality and efficient algorithms for its incremental calculation exist such as in the buckshot fractionation algorithm [23] and [24].

Other Clustering Methods:

A number of other clustering methods are known in literature; density-based method. Model-based and fuzzy clustering . Self-organization maps and even biology-inspired algorithm. An interested reader can find many surveys and book providing comprehensive information on the subject.

There are mainly two types of intrusion detection techniques, supervised learning and unsupervised learning techniques. As:

Unsupervised learning	Supervised learning
➢ The model is not provided with the correct result during the training. ➢ Can be used to cluster the input data is classes on the basis of their statistical properties only. ➢ Cluster significance and labelling. ➢ The labelling can be carried out even if the labels Are only available for a small number of objects representative of the desired cluster.	➢ Training data includes both the input and the desired results. ➢ For some examples the correct result (targets) are known and are given in input to the model during the learning process. ➢ The construction of a proper training, validation and test set (Bok) is crucial. ➢ These methods are usually fast and accurate. ➢ Have to be able to generalize: give the correct result when new data are given in input without knowing a priori the target.

Unsupervised Network Attack Detection Technique:

There are two knowledge based approaches, signature-based detection and anomaly-based detection. IDs and IPSs, and firewalls uses signature based detection. Signature base detection system can detect those attacks which it is strain to alert on. While anomaly detection uses labelled data for the creation for the normal operation traffic profiles. But as this approach requires training for profiling. Thus it becomes time consuming work. Thus this paper concentrates on tackling anomaly detection problems. There is the requirement of the analysis techniques which is not depending on knowledge, which is knowledge independent technique. Aiming at discovering knowledge independent system. New proposed algorithm is unsupervised network attack detection algorithm. The figure describing algorithm is as in figure.

Select Unsupervised Algorithms:

1. Association rules mining/ market base analysis: looks for combinations of items that occur together
2. Independent components analysis- conceptually similar to principle components analysis, but can work on variables that are not jointly normally distributed; a form of blind source/ single separation
3. K-means clustering- organizes a set of observations into clusters, where observation in a group cluster closely around a centroid / mean

4. Self-organizing maps – similar to multidimensional scaling, task a high dimensional problem and translate it into low dimensional space so it can be visualized; uses neural network to process data

Unsupervised learning- input and output are known, finds useful patterns

Clustering	**Associations / sequences**
➢ Exploratory data analysis ➢ Reveals natural groups within a data set ➢ Distance measure : no prior knowledge about groups or characteristics ➢ Not always an end in itself **Customer segmentations**	➢ Finds things that occur together ➢ Association can exist between any of the attributes ➢ Discover association rules in time- oriented data ➢ Find the sequence or order of the events **Market based analysis, next logical Purchase**

There are two types of intrusion detection techniques, supervised learning and unsupervised learning techniques. The following figure showing working difference between supervised and unsupervised learning techniques as in following fig. 1.

Fig. 1.WorkingDifference between Supervised and Unsupervised Learning Technique in Data Mining for Intrusion Detection

Initially traffic is captured and packet are analysed by aggregating them in multi flow, on the top of these flow. Different time series is built. And anomalous change is defined by change-detection algorithm based on time-series analysis.

Literature Survey Is Based On Clustering and Outlier Detection Techniques :

Autonomous network security using unsupervised detection of network attacks will work in the different way. It is completely autonomous that is without any kind of celebration or previous knowledge, if it is plugged in monitoring system, it starts to work. Second advantage is that signatures build by the system are compact and easy which can characterize attack in effective way. Third and most important advantage is that it combines the robust clustering techniques such that many clustering problem are avoided.

Methodology

Network Historical and Log Data

The network historical and log data are the network activity data. These data are passively monitored, scanned and collected through the various monitoring mechanisms.[6][7][8][9][10][11][12] are explored as follows**:**

KISMET:

kismet is a scanning tool that uses the 802.11 wireless detectors and permits card based passive monitoring (RF-MON) to sniff any 802.11x standard network it displays ARP (address resolutions protocol) and DHCP (dynamic host configuration protocol) traffic to save file in the file format of Wires hark and TCPDump and display level of activity at same different channels. It decodes and measures the real-time traffic signals. Hackers mostly use the KISMET, since it can be used in any communication network it helps to detect the intrusions. It runs an MAC and LINUX the platforms [13] [14].

SNOOP:

Sun micro systems developed a common intrusion detection sniffing tool ‘snoop’ to function with Solaris plat form. it adopts single and multiline format to display the results it sniff IPv4 and IPv6 network packets. This tool is similar to TCP Dump in displaying and formatting of the files. Snoop is considerably good than TCP Dump because its user friendly interfaces [15] [16].

WIRESHARK:

The **GERALD COMBS** developed first public packet sniffing tool “Wire shark” earlier known as Ethereal it is an open source packet sniffer and analyser and licensed by GNU GPL (general public license).it works with the FreeBSD, UNIX, Linux, Solaris, and open BSD and window platforms [17].it is user friendly to capture, filter and analyze packets. This tool is very flexible since its log files are in different format [14].

TCP Dump:

The Lawrence Berkeley national laboratory developed the TCP Dump open source network scanning and repair tool for TCP/IP packet network in 1990.the user intercepts captures and monitors TCP/IP packets during transmission in a network. it works with UNIX, Linux, Solaris, BSD (Berkeley software distribution), MAC and window platforms. it uses the command line to capture and filter log based on certain rules. These log files are not in understandable format [20] [21] [22].

Conclusion:

The completely unsupervised algorithm for detection of network attacks has many interesting advantages with respect to previous proposals. Exclusively unlabelled data is used for detection and characterization of network attacks. it does not depend or assume any signature, model, or data distribution. Thus new previously unseen attack can be detected, without using statistical learning. Robustness is removed by combining the notions of sub-space clustering and multiple evidence accumulation. The algorithm avoids the lack of robustness off general clustering approaches, improving the power of discrimination between normal-operation and anomalous traffic. Finally, we introduce the security schemes in the data mining that can help to protect the data from intrusion with the help of unsupervised learning detection technique. During the survey, we also find some points that can be further explored in future. Such as finding some effective security solutions and protecting the data mining based data using modified unsupervised or clustering techniques and detection of attacks with efficient intrusion detection mechanism. We can also explore much more in this research area.

Refrences:

1. M.Schafer. V.Lenders, and I.Martinovic.”Experimental analysis of attacks on next generation air traffic communication.” In Applied Cryptography and Network security. 2013. pp.253-271.
2. N.meng, .J.wang, E.Kodama, and T.Takata,”Reducing Data Leakage possibility resulted from eavesdropping in wireless sensor network.” International journal of space-based and situated computing, vol.3.no.1 pp.55-56, 2013
3. T.denning, T.Kohno, and H.M.Levy, “Computer Security and the Modern home.” Communication of the ACM, vol.56.no.1, pp.94-103.2013.
4. T.H.Lin, C-Y.Lin, and T.Hwang,”Man-In-The-Middle Attack on ‘Quantum Dialogue with Authentication Based on Bell States’.”International Journal of Theoretical Physics, pp.1-5, 3013.
5. Z.Tan, P.Nanda, R.P.Liu, A.Jamdagni, and X.He.” A system for Denial-of-Service Attack Detection Based on Multivariate Correlation Analysis.”IEEE Transactions on Parallel and Distributed Systems. 99. No.I.P.I, 2013.

6. A.M. Rajeswari, G. V. Aishwarya, V. A. Nachammai, and C. Deisy, “temporal outlier detection on quantities data using unexpectedness measure.” In Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on.2012, pp.420-424.
7. G.Han, J.Jiang, W.Shen, L.Shu, and I.Rodrigues. “IDSEP: a novel intrusion detection scheme based on energy prediction in cluster-based wireless sensor networks.”IET information security. Vol.7, no.2, pp.97-105, 2013.
8. H.zhao and y.shi.”Detecting covert channels in computer networks based on chaos theory.” 2013.
9. G.H. Tu, C.Peng, H.Wang, C.-Y.Li, and S.Lu, “HowVoice call affect data in operational LTE networks.” 2013.
10. B. G. Gohil, R. K. Pathak, and A. A. Patel, “Federated network security administration framework.”2013.
11. C. Thomas and n. Balakrishanan, “issues and challenges in intrusion detection with skewed network traffic.” 2013.
12. G. Ruig Utges, “vulnerability assessment of distributed system.” B. s. thesis, 2013.
13. E. G. Morgan, M. G. Sheen, F. Alizadeh-Shabdiz, and R. K. Jones, Continuous data optimization of model access points in positioning systems. 2013.
14. F. Li, M. Li, R. Lu, H. Wu, m. Claypool, and r. Kinicki,”tools and techniques for measurement of ieee802.11 wireless networks.” In modelling and optimization in mobile, ad hoc and wireless networks, 2006, pp.1-8
15. D. Dasgupta, and h. Brian, “Mobile security agents for network traffic analysis” in DARPA information survivability conference and exposition 11. 2001, DICEX’01. Proceeding, 2001, vol. 2, pp.332-340.
16. P. Li, C. Li, and T. Mohammed. “Building a repository of network traffic capture for information assurance education.” Journal of computing science in colleges. Vol.24, no.3, pp. 99-105, 2009.
17. Rakesh Shrestha, 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 network and applications, pp. 647-665.
18. Shaik Akbar, Dr. K. Nageshwara Rao, and Dr. J. A. Chandulal, “intrusion detection system methodologies based on data analysis.” International journal of computer applications (0975-8887) vol. 5, no. 2, august 2010, pp.10-20.
19. Abhinav Shrivastav, 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.
20. J. Therdphapiyank, and K. Piromsopa,”an analysis of suitable parameters for efficiently applying K-means clustering to large TCPDump dataset using Hadoop framework.” In electrical engineering/electronics. Computer, telecommunications and information technologies (ECTI-CON). 2013 10th international conference on 2013, pp.1-6.
21. N. T. Anh, and R. Shorey,”Network sniffing tools for WLANs: merits and limitations.”In personal wireless communications. 2005. ICPWC 2005. 2005 IEEE International conference on 2005, pp. 389-393.
22. F. Fuents and D. C. Kar, “Ethereal vs. TCPDump: a comparative study on packet sniffing tool for educational purpose.” Journal of computing sciences in colleges. vol. 20, no. 4, pp.169-176, 2005.
23. Preetee K.Karmore, Smita m. Nikhi, “Dectecting 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.
24. Prakash Ranganathan, Juan Li. Kendall Nygard, “A Multiagent system using associate rule mining (ARM). A collaborative filtering approach.”IEEE 2010, vol. 7, pp. 574-578.