RESEARCH ARTICLE OPEN ACCESS

Data Mining Technology for Efficient Network Security Management

Ankit Naik [1], S.W. Ahmad [2]
Student [1], Assistant Professor [2]
Department of Computer Science and Engineering
PRMIT & R, Badnera
Maharashtra – India

ABSTRACT

Several Network systems are suffering from various security threats including network worms, large scale network attacks, etc, and network security situation awareness is an effective way for solve these problems. The general process is to perceive the network security events happened in a certain time period and cyberspace environment, synthetically manipulate the security data, analyse the attack behaviours systems suffered, provide the global view of network security, and assess the whole security situation and predict the future security trends of the network.

Keywords:- Network Security, Data Mining, Network Security Situation Awareness(NSSA), Intrusion Detection System, IDS.

I. INTRODUCTION

Network security has been the eternal hot research spot, and it has undergone three phases: defence, detection and fault. However, there are still some security problems left, such as the complicated structure, and the kittle network attacks, so the network security issues are becoming more and more austere. The existed professional network security means, like IDS, Firewall and VDS cannot reflect the security status of the network. After that, Tim Bass, the American outstanding network security expert, proposes the concept of network security situational awareness (NSSA), and establishes the framework of network situation awareness, which aims to solve the existed network security problems from a new point of view[1].

II. NSSA MODEL

The realization of situational awareness is divided into three layers: perception of elements in current situation, comprehension of current situation and projection of future status. In order to evaluate the network security status of a large scale network and analyze the influence on network security of attacks or combination of them, a layered network security situational awareness realization model is proposed, shown in Figure1. The network security situation is determined by the network services, at the same time, the network services are influenced by the kinds of anomaly in the network systems. [2]

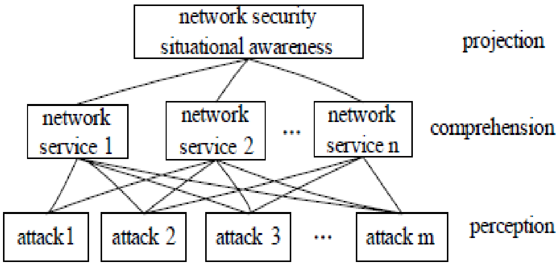

Fig 1 NSSA Realization Model

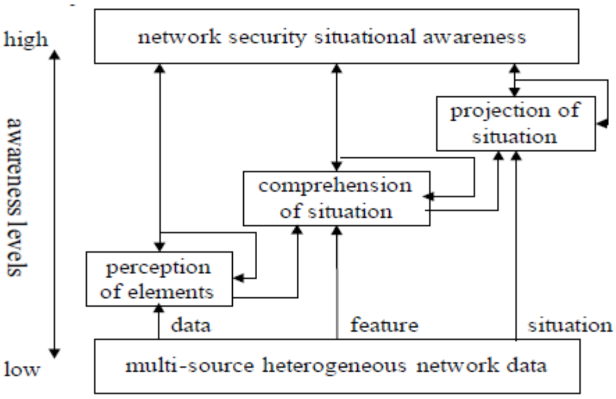
high
awareness levels
low
network security situational awareness
projection of situation
comprehension of situation
perception of elements
data
feature
situation
multi-source heterogeneous network data

Fig 2 Hierarchial NSSA Model

III. DATA MINING. WHAT IS IT?

Data mining (DM), also called Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using association rules [see fig 3]. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition. Here are a few specific things that data mining might contribute to an intrusion detection project:[3]

- Remove normal activity from alarm data to allow analysts to focus on real attacks

- Identify false alarm generators and ”bad” sensor signatures

- Find anomalous activity that uncovers a real attack

- Identify long, ongoing patterns (different IP address, same activity)

To accomplish these tasks, data miners employ one or more of the following techniques:

- Data summarization with statistics, including finding outliers

- Visualization: presenting a graphical summary of the data
- Clustering of the data into natural categories

- Association rule discovery: defining normal activity and enabling the discovery of anomalies

- Classification: predicting the category to which a particular record belongs

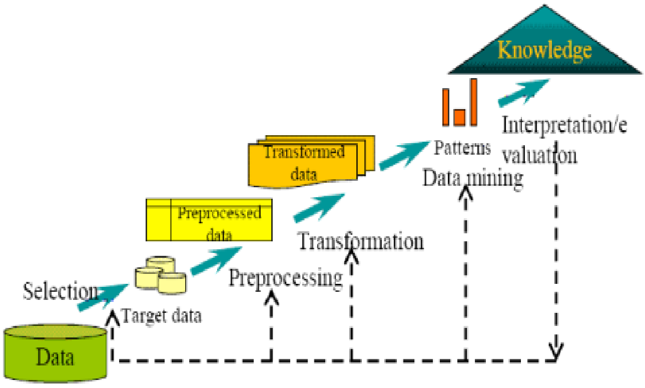

Fig 3 Transition from raw data to vulnerable knowledge

IV. DATA MINING AND INTRUSION DETECTION

Data mining is becoming an important component in intrusion detection system. Different data mining approaches like classification, clustering etc are frequently used to analyze network data to gain intrusion related knowledge. This section will elaborate on several of these data mining techniques and will describe how they are used in the context of intrusion detection.[7]

A. *Classification analysis*

[8]Classification algorithms can be used for both misuse and anomaly detections. In misuse detection, network traffic data are collected and labelled as “normal” or “intrusion”. This labelled dataset is used as a training data to learn classifiers of different types (e.g., SVM,NN,NB, or ID3) which can be used to detect known intrusions. In anomaly detection, the normal behaviour model is learned from the training dataset that are known to be “normal” using learning algorithms.

Classification models can be built using a wide variety of algorithms. Classification categorizes the data records in a predetermined set of classes used as attribute to label each record; distinguishing elements belonging to the normal or abnormal class. This technique has been popular to detect individual attacks but has to be applied with complementary fine-tuning techniques to reduce its demonstrated high false positives rate.

B. *Clustering*

[8]Clustering is the process of labeling data and assigning it into groups. Clustering algorithms can group new data instances into similar groups. These groups can be used to increase the performance of existing classifiers. High quality clusters can also assist human expert with labeling. A cluster is 100% pure if it contains only data instances from one category. Clustering techniques can be categorized into the following classes: pairwise clustering and central clustering. Pairwise clustering (i.e., similaritybased clustering) unifies similar data instances based on a data-pairwise distance measure. On the other hand, Central clustering, also called centroid-based or model-based clustering, models each cluster by its "centroid". In terms of runtime complexity, centroid-based clustering algorithms are more efficient than similarity-based clustering algorithms.

C. *Association and correlation analysis*

The main objective of association rule analysis is to discover association relationships between specific values of features in large datasets. This helps discover hidden patterns and has a wide variety of applications in business and research. Association rules can help select discriminating attributes that are useful for intrusion detection. It can be applied to find relationships between system attributes describing network data. New attributes derived from aggregated data may also be helpful, such as summary counts of traffic matching a particular pattern.

D. *Stream data analysis*

Intrusions and malicious attacks are of dynamic nature. Moreover, data streams may help detect intrusions in the sense that an event may be normal on its own. Thus, it is necessary to perform intrusion detection in data stream, real-time environment. This helps identify sequences of events that are frequently encountered together, find sequential patterns, and identify outliers. Other data mining methods for finding evolving clusters and building dynamic classification models in data streams can be applied for these purposes.

E. *Distributed data mining:*

Intruders can work from several different locations and attack many different destinations. Distributed data mining methods may be utilized to analyze network data from several network locations, this helps detect distributed attacks and prevent attackers in different places from harming our data and resources. Visualization and querying tools: Visualization data mining tools that include features to view classes, associations, clusters, and outliers can be used for viewing any anomalous patterns detected. Graphical user interface associated with these tools allows security analysts to understand intrusion detection results, evaluate IDS performance and decide on future enhancements for the system.

V. DATA MINING ALGORITHMS TO IMPLEMENT INTRUSION DETCTION SYSTEM

Most Popular Data Mining Algorithms for IDs

A. *Bayes Classifier*

Bayesian network is a model that encodes probabilistic relationships among variables of interest. This technique is generally used for intrusion detection in combination with statistical schemes, a procedure that yields several advantages, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data. However, a serious disadvantage of using Bayesian networks is that their results are similar to those derived from threshold-based systems, while considerably higher computational effort is required.

B. *K-Nearest Neighbour*

K-Nearest Neighbour (k-NN) is instance based learning for classifying objects based on closest training examples in the feature space. It is a type of lazy learning where the function is only approximated locally and all computation s deferred until classification. The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbors. If k=1, then the object is simply assigned to the class of its nearest neighbor. The k-NN algorithm uses all labeled training instances as a model of the target function. During the classification phase, k-NN uses a similarity-based search strategy to determine a locally optimal hypothesis function. Test instances are compared to the stored instances and are assigned the same class label as the k most similar stored instances. Generally it is used for intrusion detection in combination with statistical schemes (anomaly detection).

C. *Decision Tree*

Decision tree is a predictive modeling technique most often used for classification in data mining. The Classification algorithm is inductively learned to construct a model from the preclassified data set. Each data item is defined by values of the attributes. Classification may be viewed as mapping from a set of attributes to a particular class. The Decision tree classifies the given data item using the values of its attributes. The decision tree is initially constructed from a set of pre-classified data. The main approach is to select the attributes, which best divides the data items into their classes. According to the values of these attributes the data items are partitioned. This process is recursively applied to each partitioned subset of the data items.

The process terminates when all the data items in current subset belongs to the same class. A node of a decision tree specifies an attribute by which the data is to be partitioned. Each node has a number of edges, which are labeled according to a possible value of the attribute in the parent node. An edge connects either two nodes or a node and a leaf. Leaves are labeled with a decision value for categorization of the data. Induction of the decision tree uses the training data, which is described in terms of the attributes. The main problem here is deciding the attribute, which will best partition the data into various classes. To classify an unknown object, one starts at the root of the decision tree and follows the branch indicated by the outcome of each test until a leaf node is reached. The name of the class at the leaf node is the resulting classification. Decision trees can be used as a misuse intrusion detection as they can learn a model based on the training data and can predict the future data as one of the attack types or normal based on the learned model. Decision trees work well with large data sets. This is important as large amounts of data flow across computer networks. The high performance of Decision trees makes them useful in real-time intrusion detection. Decision trees construct easily interpretable models, which is useful for a security officer to inspect and edit. These models can also be used in the rule-based models with minimum processing. Generalization accuracy of decision trees is another useful property for intrusion detection model. There will always be some new attacks on the system which are small variations of known attacks after the intrusion detection models are built. The ability to detect these new intrusions is possible due to the generalization accuracy of decision trees.

D. *Neural Network (NN)*

Neural networks have been used both in anomaly intrusion detection as well as in misuse intrusion detection. For anomaly intrusion detection, neural networks were modeled to learn the typical characteristics of system users and identify statistically significant variations from the user's established behavior. In misuse intrusion detection the neural network would receive data from the network stream and analyze the information for instances of misuse. A NN for misuse detection is implemented [5] in two ways. The first approach incorporates the neural network component into an existing or modified expert system. This method uses the neural network to filter the incoming data for suspicious events and forward them to the expert system.

This improves the effectiveness of the detection system. The second approach uses the neural network as a stand alone misuse detection system. In this method, the neural network would receive data from the network stream and analyze it for misuse intrusion. There are several advantages to this approach. It has the ability to learn the characteristics of misuse attacks and identify instances that are unlike any which have been observed before by the network. It has high degree of accuracy to recognize known suspicious events. Generally, it is used to learn complex non linear input-output relationships.

E. *Support Vector Machine*

Support Vector Machines [6] have been proposed as a novel technique for intrusion detection.An SVM maps input (real-valued) feature vectors into a higher-dimensiona feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization [11]. Structural risk minimization seeks to find a hypothesis h for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set. Computing the hyper plane to separate the data points i.e. training an SVM leads to a quadratic optimization problem. SVM uses a linear separating hyper plane to create a classifier but all the problems cannot be separated linearly in the original input space. SVM uses a feature called kernel to solve this problem. The Kernel transforms linear algorithms into nonlinear ones via a map into feature spaces. There are many kernel functions; including polynomial, radial basis functions, two layer sigmoid neural nets etc. The user may provide one of these functions at the time of training the classifier, which selects support vectors along the surface of this function.

SVMs classify data by using these support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. Computing the hyper plane to separate the data points i.e. training a SVM leads to quadratic optimization problem. SVM uses a feature called kernel to solve this problem. Kernel transforms linear algorithms into nonlinear ones via a map into feature spaces. There are many kernel functions; some of them are Polynomial, radial basis functions, two layer sigmoid neural nets etc. The user may provide one of these functions at the time of training classifier, which selects support vectors along the surface of this function. SVMs classify data by using these support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. The implementation of SVM intrusion detection system has two phases: training and testing. SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification. The main disadvantage is SVM can only handle binary-class classification whereas intrusion detection requires multi-class classification.

VI. CONCLUSION

In this paper, we describe different data mining technique and their usefulness in the context of Network Security and intrusion detection system. This paper also provides the description of the current Intrusion Detection Systems(IDS) that make use of data mining for detecting intrusion.

REFERENCES

[1] Tim Bass, “Multisensor Data Fusion for Next Generation Distributed Intrusion Detection Systems”, Proceedings of 1999 IRIS National Symposium on Sensor and Data Fusion, University of The Johns Hopkins, America, pp. 1-6, 1999.

[2] Quantification Of Network Security Situational Awareness Based On Evolutionary Neural Network.

[3] Data Mining Techniques for (Network) Intrusion Detection Systems Theodoros Lappas and Konstantinos Pelechrinis

[4] Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems Chang-Tien Lu, Arnold P. Boedihardjo, Prajwal Manalwar

[5] J. Cannady. Artificial Neural Networks for Misuse Detection. National Information Systems Security Conference, 1998.

[6] S. Mukkamala, G. Janoski, A. Sung. Intrusion Detection Using Neural Networks and Support Vector Machines. Proceedings of IEEE International Joint Conference n Neural Networks, pp.1702-1707, 2002

[7] Luan, J. Data Mining and Its Applications in Higher Education. – New Directions for Institutional Research, Special Issue Titled Knowledge Management: Building a Competitive Advantage in Higher Education, Vol. 2002, 2002, Issue 113, 17-36

[8] Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, “The Survey of Data Mining Applications And Feature Scope”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), June 2012

[9] http://www.wisegeek.com/what-are-the -different-types-of-data-mining-analysis.htm

[10] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, "Comparison the various clustering algorithms of weka tools", International Journal of Emerging Technology and Advanced Engineering, (ISSN 2250-2459, Volume 2, Issue 5, May 2012).