COMPARATIVE STUDY OF DATA MINING TECHNIQUES IN INTRUSION DECTECTION

Mousmi A. Chaurasia
Professor, MJCET, Hyderabad, Telangana.

Abstract

In these days an increasing number of public and commercial services are used through the Internet, so that security of information becomes more important issue in the society information Intrusion Detection System (IDS) used against attacks for protected to the Computer networks. On another way, some data mining techniques also contribute to intrusion detection. Some data mining techniques used for intrusion detection can be classified into two classes: misuse intrusion detection and anomaly intrusion detection. Misuse always refers to known attacks and harmful activities that exploit the known sensitivity of the system. Anomaly generally means a generally activity that is able to indicate an intrusion. The key ideas are to use data mining techniques to discover consistent and useful patterns of system features that describe program and user behavior, and use the set of relevant system features to compute (inductively learned) classifiers that can recognize anomalies and known intrusions. In this paper, comparison made between various data mining techniques for intrusion detection. Our work provide an overview on data mining and soft computing techniques such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Trees, Bayesian Finally, a discussion of the future technologies and methodologies which promise to enhance the ability of computer systems to detect intrusion is provided and current research challenges are pointed out in the field of intrusion detection system.

Keywords: intrusion detection, data mining, ANN, SVM

I. 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 and networking sources [1]. Intrusion prevention techniques, such as user authentication and information protection have been used to protect computer systems as a first line of defense. Intrusion prevention alone is not sufficient because as systems become ever more complex, there are always exploitable weaknesses in the systems due to design and programming errors. Now a day, intrusion detection is one of the high priority tasks for network administrators and security professionals.

As network based computer systems play increasingly vital roles in modern society, they have become intrusion detection systems provide following three 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 their 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 a system resource that ensures that it is accessible and usable upon demand by an authorized system user.

Any intrusion detection system has some inherent requirements. Its prime purpose is to detect as many attacks as possible with minimum number of false alarms, i.e. the system must be accurate in detecting attacks. However, an accurate system that cannot handle large amount of network traffic and is slow in decision making will not fulfill the purpose of an intrusion detection system (IDs). Data mining techniques like data reduction, data classification, features selection techniques play an important role in IDS. Intrusion detection is therefore needed as another wall to protect computer systems. The elements central to intrusion detection are: *resources* to be protected in a target system, i.e., user accounts, file systems, system kernels, etc; *models* that characterize the "normal" or "legitimate" behavior of these resources; *techniques* that compare the actual system activities with the established models, and identify those that are "abnormal" or "intrusive".

Many researchers have proposed and implemented different models which define different measures of system behavior, with an ad hoc presumption that normalcy and anomaly (or illegitimacy) will be accurately manifested in the chosen set of system features that are modeled and measured. Intrusion detection techniques can be categorized into *misuse detection*, which uses patterns of well-known attacks or weak spots of the system to identify intrusions; and *anomaly detection*, which tries to determine whether deviation from the established normal usage patterns can be flagged as intrusions.

Misuse detection systems, for example [2] and STAT [3], encode and match the sequence of "signature actions" (e.g., change the ownership of a file) of known intrusion scenarios. The main shortcomings of such systems are: known intrusion patterns have to be hand-coded into the system; they are unable to detect any future (unknown) intrusions that have no matched patterns stored in the system.

Anomaly detection (sub)systems, such as IDES [4], establish normal usage patterns (profiles) using statistical measures on system features, for example, the CPU and I/O activities by a particular user or program. The main difficulties of these systems are: intuition and experience is relied upon in selecting the system features, which can vary greatly among different computing environments; some intrusions can only be detected by studying the sequential interrelation between events because each event alone may fit the profiles.

Table 1: a comparison between the two types of intrusion detection

	Misuse Detection	Anomaly Detection
Characteristics	use patterns of well-known attacks (signatures) to identify intrusions, any match with signatures is reported as a possible attack	use deviation from normal usage patterns to identify intrusions, any significant deviations from the expected behaviour are reported as possible attacks
Drawbacks	- False negatives - Unable to detect new attacks - Need signatures update - Known attacks has to be hand-coded - Overwhelming security analysts	-False positives. - Selecting the right set of system features to be measured is ad hoc and based on experience - Has to study sequential interrelation between transactions - Overwhelming security analysts

From the above discussion, we conclude that traditional IDS face many limitations. This has led to an increased interest in improving current IDS. Applying Data Mining (DM) techniques such as classification, clustering, association rules, etc, on network traffic data is a promising solution that helps improve IDS [11-19].

As there are many number of ID techniques using data mining techniques, the unknown

technique and system could be thought of as a baseline for future prospect. As a result, the purpose of this paper is to review related papers of using data mining for intrusion detection. The contribution of this research paper is to provide a comparison of IDS in terms of data mining IDS techniques used for future research directions. This paper is organized as follows. Section 2 overviews the data mining techniques for intrusion detection. Section 3 compares related work of IDS. Section 4 discusses about the future work and the conclusion is also provided.

II. LITERATURE REVIEW

Intrusion detection is the process of monitoring and analyzing the data and events occurring in a computer and/or network system in order to detect attacks, vulnerabilities and other security problems [5]. IDS can be classified according to data sources into: host-based detection and network-based detection. In host-based detection, data files and OS processes of the host are directly monitored to determine exactly which host resources are the targets of a particular attack. In contrast, network-based detection systems monitor network traffic data using a set of sensors attached to the network to capture any malicious activities. Networks security problems can vary widely and can affect different security requirements including authentication, integrity, authorization, and availability. Intruders can cause different types of attacks such as Denial of Services (DoS), scan, compromises, and worms and viruses [6,7]

A. Data Mining used in IDS

Much number of data mining techniques can be used in intrusion detection, each with its own specific advantage. The following lists some of the techniques and the motives for which they may be employed. Classification: Creates a classification of tuples. It could be used to detect individual attacks, but as described by previous sample experiments in the literature indication it is produce a high false alarm rate. This problem may be reduced by applying fine-tuning techniques such as boosting. Association: Describes relationships within tuples. Detection of irregularities may occur when many tuples exhibit previously unseen relationships. Grouping: Groups tuples that exhibit similar properties according to pre-described metrics. It can be used for general analysis similar to categorization, or for detecting outliers that may or may not represent attacks.

B. Models

1. Artificial Neural Network (ANN):

Artificial Neural Network (ANN) is relatively crude electronic models based on the neural structure of the brain. The brain basically learns from his experience. This is natural proof that some problems that are beyond the scope and range of current computers are indeed solvable by small energy efficient packages. This brain modeling a technical way to develop machine solutions. This new arrival approach to computing also provides a more graceful degradation during system overload than its more habitual counterparts.

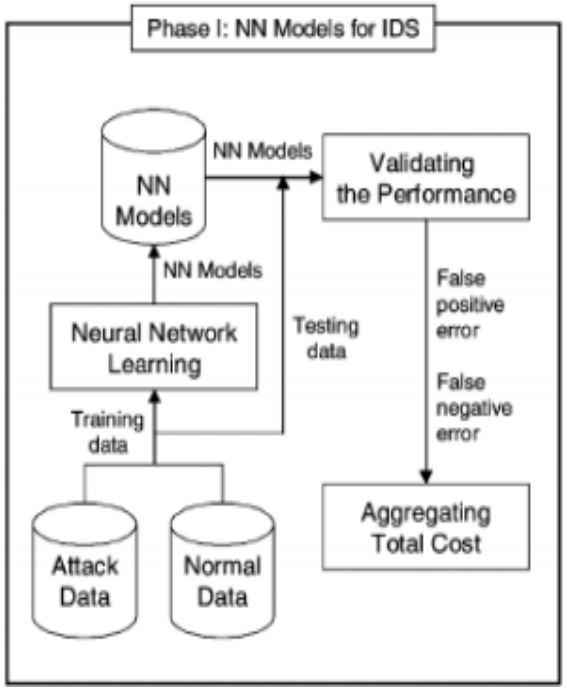

Figure 1, Shows a neural network model used for IDS. [8]

A neural network is an interrelated group of artificial neurons that uses a mathematical model or computational model for information processing based on a connection approach to computation. A neural network could not contains domain knowledge in the beginning, but it can be supervised to make decisions by mapping example pairs of input data into example output vectors, and estimating its weights so that it maps each input example vector into the corresponding output example vector approx.

2. Support vector machines (SVM)

The SVM approach converts data into a feature space F that usually has a large dimension. This is interesting to note that SVM generalization depends on the geometrical properties of the supervised data, not on the dimensions of the input space [9]. Detail information of SVM can be found in [10].

3. Bayes Classifier

A 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.

4. 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).

5. 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. 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.

Table 1 shows a general comparison of different classifiers. There are various approaches[11-19] to implement an intrusion detection system based on its type and mode of deployment. Each of the approaches to implement an intrusion detection system has its own advantages and disadvantages. This is apparent from the discussion of comparison among the various methods. Thus it is difficult to choose a particular method to implement an intrusion detection system over the other.

Classifier	Method	Parameters	Advantages	Disadvantages
Support Vector Machine	A support vector machine constructs a hyper plane or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks.	The effectiveness of SVM lies in the selection of kernel and soft margin parameters. For kernels, different pairs of (C, γ) values are tried and the one with the best cross-validation accuracy is picked. Trying exponentially growing sequences of C is a practical method to identify good parameters.	1. Highly Accurate 2. Able to model complex nonlinear decision boundaries 3. Less prone to over fitting than other methods	1. High algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks. 2. The choice of the kernel is difficult 3. The speed both in training and testing is slow.
K Nearest Neighbour	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 neighbours (*k* is a positive integer). If *k* = 1, then the object is simply assigned to the class of its nearest neighbour.	Two parameters are considered to optimize the performance of the kNN, the number k of nearest neighbour and the feature space transformation.	1. Analytically tractable. 2. Simple in implementation 3. Uses local information, which can yield highly adaptive behaviour 4. Lends itself very easily to parallel implementations	1. Large storage requirements. 2. Highly susceptible to the curse of dimensionality. 3. Slow in classifying test tuples.
Artificial Neural Network	An ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.	ANN uses the cost function *C* is an important concept in learning, as it is a measure of how far away a particular solution is from an optimal solution to the problem to be solved.	1. Requires less formal statistical training. 2. Able to implicitly detect complex nonlinear relationships between dependent and independent variables. 3. High tolerance to noisy data. 4. Availability of multiple training algorithms.	1. "Black box" nature. 2. Greater computational burden. 3. Proneness to over fitting. 4. Requires long training time.
Bayesian Method	Based on the rule, using the joint probabilities of sample observations and classes, the algorithm attempts to estimate the conditional probabilities of classes given an observation.	In Bayes, all model parameters (*i.e.*, class priors and feature probability distributions) can be approximated with relative frequencies from the training set.	1. Naïve Bayesian classifier simplifies the computations. 2. Exhibit high accuracy and speed when applied to large databases.	1 The assumptions made in class conditional independence. 2. Lack of available probability data.
Decision Tree	Decision tree builds a binary classification tree. Each node corresponds to a binary predicate on one attribute; one branch corresponds to the positive instances of the predicate and the other to the negative instances.	Decision Tree Induction uses parameters like a set of candidate attributes and an attribute selection method.	1. Construction does not require any domain knowledge. 2. Can handle high dimensional data. 3. Representation is easy to understand. 4. Able to process both numerical and categorical data.	1. Output attribute must be categorical. 2. Limited to one output attribute. 3. Decision tree algorithms are unstable. 4. Trees created from numeric datasets can be complex.

C. Research Challenges

Following are the research challenges of the existing intrusion detection classification problem using data mining technique:

1. The final decision must be wrong if the output of selected classifier is wrong.
2. The trained classifier may not be complex enough to handle the problem

CONCLUSION

The security of computer networks plays a planning role in modern computer system. Detection of intrusion attacks is the most important issue in computer network security. This paper draws the conclusions on the basis of implementations performed using various data mining algorithms. Combining more than one data mining algorithms may be used to

remove disadvantages of one another. Thus a combining approach has to be made while selecting a mode to implement intrusion detection system. Combining a number of trained classifiers lead to a better performance than any single classifier.

References

[1] Amoroso EG (1999) Intrusion detection: an introduction to internet surveillance, correlation, trace back, traps, and response. Intrusion.Net Books, NJ

[2] S. Kumar and E. H. Spafford.
A software architecture to support misuse intrusion detection.
In Proceedings of the 18th National Information Security Conference, pages 194-204, 1995.

[3] K. Ilgun, R. A. Kemmerer, and P. A. Porras.
State transition analysis: A rule-based intrusion detection approach.
IEEE Transactions on Software Engineering, 21(3):181-199, March 1995.

[4] T. Lunt, A. Tamaru, F. Gilham, R. Jagannathan, P. Neumann, H. Javitz, A. Valdes, and T. Garvey.
A real-time intrusion detection expert system (IDES) - final technical report. Technical report, Computer Science Laboratory, SRI International, Menlo Park, California, February 1992.

[5] Jiawei Han and. Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kufmann, 2nd edition 2006, 3rd edition 2011.

[6] S.J. Stolfo, W. Lee. P. Chan, W. Fan and E. Eskin, "Data Mining – based Intrusion Detector: An overview of the Columbia IDS Project" ACM SIGMOD Records vol. 30, Issue 4, 2001.

[7] W. Lee and S.J. Stolfo, "Data Mining Approaches for Intrusion Detection" 7th USENIX Security Symposium, Texas, 1998.

[8] Joo, D., Hong, T., and Han, I. "The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors." Expert Systems with Applications 25, 69–75 2000.

[9] Joachims, T. "SVM light is an implementation of support vector machines (SVMs) in C". University of Dortmund" Collaborative Research Center on Complexity Reduction in Multivariate Data (SFB475), http://ais.gmd.de/,thorsten/svm_light 2000.

[10] Vaprik, V. "Statistics learning theory." John Wiley, New York. (1998)

[11] Susan M. Bridges , Rayford B. Vaughn, "Data mining and genetic algorithms applied to intrusion detection", In Proceedings of the National Information Systems Security Conference, 2000.

[12] Jiawei Han and. Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kufmann, 2nd edition 2006, 3rd edition 2011.

[13] S.J. Stolfo, W. Lee. P. Chan, W. Fan and E. Eskin, "Data Mining – based Intrusion Detector: An overview of the Columbia IDS Project" ACM SIGMOD Records vol. 30, Issue 4, 2001.

[14] W. Lee and S.J. Stolfo, "Data Mining Approaches for Intrusion Detection" 7th USENIX Security Symposium, Texas, 1998.

[15] S. Terry Brugger, "Data Mining Methods for Network Intrusion detection", University of California, Davis, 2004. http://www.mendeley.com/research/data-mining-methods-for-network-intrusion-detection/

[16] T. Lappas and K. Pelechrinis, "Data Mining Techniques for (Network) Intrusion Detection Systems" http://atl-svn.assembla.com/svn/odinIDS/Egio/artigos/datamining/dataIDS.pdf

[17] P. Dokas, L. Ertoz, V. Kumar, A. Lazaevic. J. Srivastava, and P. Tan, "Data Mining for Network Intrusion Detection", 2002,
http://minds.cs.umn.edu/papers/nsf_ngdm_2002.pdf

[18] M. Hossain "Data Mining Approaches for Intrusion Detection: Issues and Research Directions",
http://www.cse.msstate.edu/~bridges/papers/iasted.pdf

[19] A. Chauhan, G. Mishra, and G. Kumar, "Survey on Data mining Techniques in Intrusion Detection", International journal of Scientific & Engineering Research Vol.2 Issue 7, 2011.