Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 10, October 2013, pg.145 – 153

SURVEY ARTICLE

A Survey: Network Intrusion Detection System based on Data Mining Techniques

Subaira.A.S[1], Anitha.P[2]

[1]Department of CSE & Dr. N. G. P. Institute of Technology, Coimbatore, India

[2]Department of CSE & Dr. N. G. P. Institute of Technology, Coimbatore, India

[1] subairaooty@gmail.com; [2] anitha.ngp07@gmail.com

Abstract— In spite of growing information system widely, security has remained one hard-hitting area for computers as well as networks. In information protection, Intrusion Detection System (IDS) is used to safeguard the data confidentiality, integrity and system availability from various types of attacks. Data mining is an efficient artifice that can be applied to intrusion detection to ascertain a new outline from the massive network data as well as it use to reduce the strain of the manual compilations of the normal and abnormal behaviour patterns. This work reviews the present state of data mining techniques and compares various data mining techniques used to implement an intrusion detection system such as, Support Vector Machine, Genetic Algorithm, Neural network, Fuzzy Logic, Bayesian Classifier, K- Nearest Neighbour and decision tree Algorithms by highlighting the advantages and disadvantages of each of the techniques

Keywords— Classification; Data Mining; Intrusion detection system; Anomaly Detection; Misuse Detection

I. INTRODUCTION

In the era of information society, as network-based computer systems play fundamental roles, they have become the target for intrusions by attackers and criminals. Intrusion prevention technique such as firewalls, user authentication, information protection and data encryption have failed to completely shield networks and systems behaviour from the growing and sophisticated attacks and malwares. To protect the computers and networks from various cyber-attacks and viruses the Intrusion Detection Systems (IDS) are designed. An IDS is a mechanism that monitors network or system actions for malicious activities and produces reports to a management station [1].

As a significant application area of data mining is intrusion detection based on data mining algorithms, aims to solve the troubles of analyzing enormous volumes of data [8]. IDSs build efficient clustering and classification models to distinguish normal behaviour from abnormal behaviour using data mining techniques. This study makes foundation in this field of research and exploration and implements intrusion detection model system based on data mining technology.

II. TRADITIONAL INTRUSION DETECTION

There are two types of traditional intrusion detection system

A. Anomaly Detection

It refers to detection of abnormal behaviour of host or network. It actually refers to storing features of user's usual behaviour hooked on database, and then it compares user's present behaviour with database. If any deviation occurs, then the data tested is abnormal [6]. The patterns detected are called anomalies. Anomalies are also referred to as outliers.

B. Misuse Detection

In misuse detection approach, it defines abnormal system behaviour at first, and then defines any other behaviour, as normal behaviour. It assumes that detecting abnormal behaviour at first has a simple to define model. It produce high intrusion detection rate and raise low percentage of false alarm. However, it fails in discovering the non-pre-elected attacks in the feature library, so it cannot detect the abundant new attacks [16].

IDS provide the following security functions

C. Data Confidentiality

It checks whether the information stored on a system is protected against unconstitutional access. Since systems are sometimes used to manage sensitive information, data confidentiality is often a gauge of the ability of the system to protect its data [19].

D. Data Integrity

It refers to maintaining and assuring the correctness and consistency of data over its entire life-cycle. No corruption or data loss is acknowledged either from random events or malicious activity.

E. Data Availability

The network should be tough to Denial of Service attacks.

Intrusion detection system based on sources of audit information can be divided into 3 subcategories

F. Host Based IDS

It refers to intrusion detection that takes place on a single host system. It gets audit data from host audit trails and monitors activities such as integrity of system, file changes, host based network traffics, and system logs. If there is any unlawful change or movement is detected, it alerts the user by a pop-up menu and informs the central management server. Central management server blocks the movement or a combination of the above three [17]. The judgment should be based on the strategy that is installed on the local system.

G. Network Based IDS

It is used to supervise and investigate network transfer to protect a system from network-based threats. It tries to detect malicious activities such as denial-of-service (Dos) attacks and network traffic attacks. Network based IDS includes a number of sensors to monitors packet traffic, one or more servers for network management functions, and one or more management relieves for the human interface [18].

H. Hybrid Intrusion Detection

The recent development in intrusion detection is to combine both types host-based and network-based IDS to design hybrid systems. Hybrid intrusion detection system has flexibility and it increases the security level. It combines IDS sensor locations and reports attacks are aimed at particular segments or entire network [28].

III. TYPES OF ATTACKS

A. Dos attack

A denial-of-service attack or distributed denial-of-service attack is an effort to make a computer resource out of stock to its indented users [32].This type of attack slows down the system or shut down the system so it disrupt the service and deny the legitimate authorized user. Due to this attack high network traffic occurs [15].

B. User to Root Attack (U2R)

In this type of attack, the attacker starts with user level like taking down the password, dictionary attack and finally attacker achieves root to access the system.

C. Probing

In this type of attack, an attacker examines a network to gather information or discover well-known vulnerabilities. An attacker who has a record, of which machines and services are accessible on a known network, can make use of this information to look for delicate points.

D. Remote to User Attack (R2U)

In this type of attack, an attacker has the capability to send packet to a machine over a network but does not have an account on that machine, make use of some vulnerability to achieve local access as a user of that machine.

E. Eavesdropping attack

Eavesdropping is a network layer attack consisting of capturing packets from the network transmitted by others' computers and reading the sensitive information like passwords, session tokens, or any kind of confidential information.

F. Man-In-The-Middle Attack

In this the attacker makes independent connections with the victims and relays messages between them and making them believe that they are talking directly to each other over a private connection, but the fact is that the entire conversation is controlled by the attacker.

IV. DRAWBACKS OF IDS

Intrusion Detection Systems (IDS) have become an important component in security infrastructures as they permit networks administrators to identify policy variations. These policy violations range from outside attackers trying to gain unconstitutional access to intruders abusing their access. Current IDS have a number of considerable drawbacks

*False Positives***:** A major problem is the amount of false positives IDS will produce. Developing distinctive signatures is a complicated task. It is much trickier to pick out a legitimate intrusion attempt if a signature also alerts regularly on valid network activity.

False Negatives: In some cases IDSs do not generate an alert when an intrusion is actually taking place. It simply put if a signature has not been written for a particular exploit, so there is a tremendously good chance that the IDS will not detect it.

V. DATA MINING ASSISTS IN INTRUSION DETECTION

The central theme of intrusion detection using data mining approach is to detect the security violations in information system. Data mining can process large amount of data and it discovers hidden and ignored information. To detect the intrusion, data mining consists of following processes such as classification, clustering, and regression [3]. It monitors the information system and raises alarms when security violations are founded.

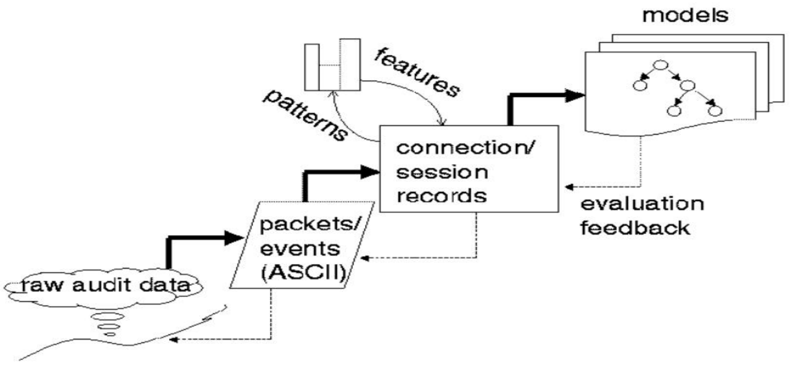

Fig. 1[30]**:** The Data Mining Process of Building ID Models

A. *Support Vector Machine (SVM)*

SVM is a learning method for the Classification and Regression analysis of both linear and nonlinear data. It uses a hypothesis space of linear functions and maps input feature vectors into a higher dimensional space all the way through some nonlinear mapping [2].SVM constructs a hyper plane or set of hyper planes only the good separation is achieved by the hyper plane. The hyper plane searching process in SVM is achieved by the leading margin [7] [13]. The related margin gives the major separation between classes. While training an SVM it creates a quadratic optimization problem [4].

In SVM the classifier is created by linear separating hyper plane but all the linear separation cannot be solved in the original input space. SVM uses a function called kernel to solve this problem. The Kernel transforms linear problem into nonlinear one by mapping into feature spaces. Radial basis function, polynomial, two layer sigmoid neural nets are the some of the kernel functions. At the time of training classifier, user may provide one of these functions, which selects support vectors along the surface of this function. The implementation of SVM tries to accomplish maximum separation between the classes [25]. Intrusion detection system has two phases: training and testing. SVMs can learn a larger set of patterns and be able to provide better classification, because the categorization difficulty 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 [11].

B. Genetic Algorithms

Genetic algorithms were initially introduced in the meadow of computational biology. After that they have been bloomed into various fields with promising result [24]. Nowadays the researchers have tried to incorporate this algorithm with IDSs. Using Genetic approach, in 1995 Giordana and Neri has proposed one intrusion detection algorithms called REGAl. The REGAL System is based on distributed genetic algorithm. REGAL is a concept learning system that learns First Order Logic multi-model concept descriptions. The learning examples are stored in relational database that are represented as relational tuples.

Gonzalez and Dasgupta [26] applied a genetic algorithm, though they were examined host based IDSs, not network based. They used the algorithm only for the Meta learning step instead of running algorithm directly on the feature set. It uses the statistical classifiers for labelled vectors. A 2-bit binary encoding methodology is used for identifying the abnormality of a particular feature, ranging from normal to abnormal. Chittur [27] used a genetic algorithm with decision tree. Decision tree is used to represent the data. They used the high detection rate that reduces the false positive rate. The false positive occurrence was minimized by utilizing human input in a feedback loop [10].

C. K-nearest Neighbour

K-Nearest Neighbour (k-NN) is a type of Lazy learning, it simply stores a given training tuple and waits until it is given a test tuple. It is an instance based learner that classifies the objects based on closet training examples in the feature space. For a given unknown tuple, a k-Nearest neighbour looks the pattern space for the k-training tuples that are closest to the unknown tuple. It is the simplest algorithm among all the machine learning algorithms. Here the object is classified by a majority vote of its neighbours. The object is simply assigned to the class of its neighbour only in the case of K=1. For a target function this algorithm uses all labelled training instances model. To obtain the optimal hypothesis function algorithm uses similarity based search. The intrusion is detected with the combination of statistical schemes. This technique is computationally expensive and requires efficient storage for implementation of parallel hardware.

D. Neural Networks

Neural Network was traditionally used to refer a network or biological neurons. In [20], IDS neural network has been used for both anomaly and misuse intrusion detection. In anomaly intrusion detection the neural networks were modelled to recognize statistically significant variations from the user’s recognized behaviour—also identify the typical characteristics of system users. In misuse intrusion detection the neural network would collect data from the network stream and analyse the data for instances of misuse [22].

In neural network the misuse intrusion detection can be implemented in two ways. The first approach incorporates the neural network component into an existing system or expert system. This method uses the neural network to sort the incoming data for suspicious events and forward them to the existing and expert system. This improves the efficiency of the detection system. The second method uses the standalone misuse detection system. This system receives data from the network stream and analyses it for misuse intrusion. 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 nonlinear input-output relationships [12].

E. Bayesian Classifier

A Bayesian Classifier provides high accuracy and speed for handling large database. In network model Bayesian classifier encodes the probabilistic relationship among the variable of interest. In intrusion detection this classifier is combined with statistical schemes to produce higher encoding interdependencies between the variables and predicting events. The graphical model of casual relationships performs learning technique. This technique is defined by two components-a directed acyclic graph and a set of conditional probability tables. Direct Acyclic Graph (DAG) represents a random variable, which may be discrete or continuous. For each variable classifier maintain one conditional probability table (CPT) and it requires higher computational effort.

F. Decision Tree

Decision tree is a classification technique in data mining for predictive models. Decision tree is a flowchart like tree structure where internal node represents a test on attribute, branch represents an outcome of the test and leaf node represents a class label. From the pre classified data set it inductively learns to construct the models. Here each data item is defined by the attribute values. Initially decision tree is constructed by set of pre-classified data. The important approach is to select the attributes, which can best divide the data items into their respective classes based on these attributes the data item is partitioned [5].

This process is iteratively applied to each partitioned subset of the data items. If all the data items in current subset belongs to the same class then the process get terminated. Each node contains the number of edges, which are labelled along with a possible value of attribute in the parent node. An edge connects either a node or two nodes. Leaves are always labelled with a decision value for classification of the data [21]. To classify an unidentified object, the process is started at the root of the decision tree and followed the branch. Decision trees can be used for misuse intrusion detection that can learn a model based on the training data and predict the future data from the various types of attacks. It works well with large data sets. Decision tree model can also be used in the rule-based techniques with minimum processing. It provides high generalization accuracy [9].

G. Fuzzy Logic

Fuzzy logic is derived from fuzzy set theory; it uses the rule based systems for classification. Fuzzy can be thought of as the application side of fuzzy set theory dealing with sound thought out real world expert values for a complex problem[29].The fuzzy related data mining techniques is used to extract the patterns behaviour. The sets of fuzzy association rules are used to mine the network audit data models and to detect the anomalous behaviour the set of fuzzy association rules are generated [30][31].The audit data and mined normal data have been compared to identify the similarity. If the similarity values are below an upper limit, an alarm raises [14].

VI. A COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES FOR INTRUSION DETECTION SYSTEM

TABLE1. GENERAL CLASSIFIER COMPARISON

Classifier	Method	Advantages	Disadvantages
Support Vector Machine	A support vector machine is a classification and regression technique it constructs a hyper plane or set of hyper planes in a high or infinite dimensional space.	1. High Accuracy. 2. Able to model complex and nonlinear decision boundaries. 3. Less prone to over fitting than other methods.	1. High algorithmic complexity and extensive memory requirement. 2. The choice of the kernel is difficult. 3. The training and testing speed is slow
Genetic Algorithm	Genetic algorithm learning examples are stored in relational database that are represented as relational tuples.	1. It solves every optimization problem. 2.It solves the problems with multiple solutions 3. Easily transferred to existing models.	1. No global optimum. 2. No constant optimization response time
K Nearest Neighbour	An object classification process is achieved by the majority vote of its neighbours. The object is being assigned to the class most common amongst its k nearest neighbours. If k = 1, then the object is simply assigned to the class of its nearby neighbour.	1. Analytically tractable. 2. Implementation task is simple. 3.Highly adaptive behaviour 4.Easy for parallel implementations	1. High storage requirements. 2. Highly susceptible to the curse of dimensionality. 3. Slow in classifying and testing tuples.
Neural Network	A Neural Network is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.	1. Requires less formal statistical training. 2. Implicitly detect the complex nonlinear relationships between dependent and independent variables. 3. Highly tolerate the noisy data. 4. Availability of multiple training algorithms.	1. Process is black box. 2. Greater computational burden. 3. Over fitting. 4. It Requires long training time.
Bayesian Method	Bayesian classifier based on the rules. It uses the joint probabilities of sample classes and observations. The algorithm tries to estimate the conditional probabilities of classes given an observation.	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 initially builds a tree with classification. Each node represents a binary predicate on one attribute, one branch represents the positive instances of the predicate and the other branch represents the negative instances.	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.
Fuzzy Logic	The fuzzy logic has been used for both anomaly and misuse intrusion detection.	1. Uses linguistic variables. 2. Allows imprecise inputs. 3.Permits fuzzy thresholds 4.Reconciles conflicting objectives 5.Rule base or fuzzy sets easily modified	1. Hard to develop a model from a fuzzy system. 2. Require more fine tuning and simulation before operational.

VII. CONCLUSIONS

In this paper, many data mining techniques have been proposed to improve the classification mechanism of Network Intrusion Detection. Different classifiers have different knowledge to solve the problem so combining more than one data mining algorithm is used to remove the demerits of one another and a number of trained classifier lead to a superior performance than any single classifier. These techniques provide better performance in Intrusion Detection accuracy rate and faster running time. To molder a complex problem into sub problems for which the solutions obtained are simpler to realize, execute, supervise and update.

REFERENCES

[1] W. Lee, S.J. Stolfo, K.W. Mok, “A data mining framework for building intrusion detection models”, in: Proceedings of IEEE Symposium on Security and Privacy, 1999, pp. 120–132.

[2] W. Feng, Q. Zhng, G. Hu, J Xiangji Huang,” Mining network data for intrusion detection through combining SVMs with ant colony networks ”Future Generation Computer Systems,2013.

[3] T. Zhang, R. Ramakrishnan, M. Livny, “BIRCH: an efficient data clustering method for very large databases”, in: Proceedings of SIGMOD, ACM, 1996, pp. 103–114.

[4] L. Khan, M. Awad, B. Thuraisingham, “A new intrusion detection system using support vector machines and hierarchical clustering”, The VLDB Journal 16(2007) 507–521

[5] X. Xu, “Adaptive intrusion detection based on machine learning: feature extraction, classifier construction and sequential pattern prediction”, Information Assurance and Security 4 (2006) 237–246.

[6] J.X. Huang, J. Miao, Ben He, “High performance query expansion using adaptive co –training”, Information Processing & Management 49 (2) (2013) 441–453.

[7] Y. Li u, X. Yu, J.X. Huang, A.” An, Combining integrated sampling with SVM ensembles for learning from imbalanced datasets”, Information Processing &Management 47 (4) (2011) 617–631.

[8] V. Vapnik, “The Nature of Statistical Learning Theory”, Springer, 1999.

[9] Marcelloni, combining “supervised and unsupervised learning for data clustering”, Neural Computing & Applications 15 (3–4) (2006)289–297.

[10] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, W.-Y. Lin, “Intrusion detection by machine learning: a review”, Expert Systems with Applications 36 (2009) 11994–12000.

[11] S.X. Wu, W. Banzhaf, “The use of computational intelligence in intrusion detection systems: a review”, Applied Soft Computing 10 (2010) 1–35.

[12] H. Brahmi, I. Brahmi, S.B. Yahia, “OMC-IDS: at the cross-roads of OLAP mining and intrusion detection”, in: Advances in Knowledge Discovery and Data Mining, in: LNCS, vol. 7302, 2012, pp. 13–24.

[13] S.-J. Horng, M.-Y. Su, Y.-H. Chen, T.-W. Kao, R.-J. Chen, J.-L. Lai, C.D. Perkasa," A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications 38 (2011) 306–313.

[14] Q. Zhang, G. Hu, W. Feng, and "Design and performance evaluation of a machine learning-based method for intrusion detection", in: Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed computing, in: Studies in Computational Intelligence, vol. 295, Springer, 2010, pp. 69–83.

[15] T.A. Longstaff, J.T. Ellis, S.V. Hernan, H.F. Lipson, R.D. McMillan, L.H. Pazente,D. Simmel, "Security of the Internet", in: F. Froehlich, A. Kent (Eds.), The Froehlich/Kent Encyclopedia of Telecommunications. Vol. 15, Marcel Derrek, 1998, pp. 231–254.

[16] S. Axelsson," Research in intrusion detection systems a survey", in: Tech. Rep.TR98-17, Chalmers University of Technology, Goteborg, Sweden, 2000.

[17] S. Freeman, J. Branch, "Host-based intrusion detection using user signatures", in: Proceedings of the Research Conference RPI., 2002.

[18] D. Marchette, "A statistical method for profiling network traffic", in: Proceedings f Workshop on Intrusion Detection and Network Monitoring, 1999,pp. 119–128.

[19] T.F. Lunt, "A survey of intrusion detection techniques", Computers and Security12 (4) (1993) 405–418.

[20] J. Ryan, M.-J. Lin, R. Miikkulainen,"Intrusion detection with neural networks", in: Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection and Task Management, 1997, pp. 92–97.

[21] H. Teng, K. Chen, S. Lu, "Security audit trail analysis using inductively generated predictive rules", in: Proceedings of the 6th Conference on Artificial Intelligence Applications, Vol. 1, 1990, pp. 24–29.

[22]] D.E. Denning," An intrusion-detection model", IEEE Transactions on Software Engineering 13 (2) (1987) 222–232.

[23] F. Monrose, A. Rubin, "Authentication via keystroke dynamics", in: Proceedings of the 4th ACM Conference on Computer and Communications Security, 1997

[24] Neri, F., "Comparing local search with respect to genetic evolution to detect intrusion in computer networks", In Proc. of the 2000 Congress on Evolutionary Computation CEC00, La Jolla, CA, pp. 238243. IEEE Press, 16-19 July, 2000.

[25] Neri, F. "Mining TCP/IP traffic for network intrusion detection", In R. L.de M'antaras and E. Plaza (Eds.), Proc. of Machine Learning: ECML\2000, 11th European Conference on Machine Learning, Volume 1810of Lecture Notes in Computer Science, Barcelona, Spain, pp. 313322.Springer, May 31- June 2, 2000.

[26] Dasgupta, D. and F. A. Gonzalez,"An intelligent decision support system for intrusion detection and response", In Proc. of International Workshop on Mathematical Methods, Models and Architectures for Computer Networks Security (MMM-ACNS), St.Petersburg. Springer-Verlag, 21-23 May, 2001

[27] Chittur, A., "Model generation for an intrusion detection system using genetic algorithms", High School Honors Thesis, Ossining High School. In cooperation with Columbia Univ, 2001.

[28] Crosbie, M. and E. H. Spafford,"Active defense of a computer system agents", Technical Report CSD-TR-95-008, Purdue Univ. West Lafayette, IN, 15 February 1995.

[29] G. J. Klir,"Fuzzy arithmetic with requisite constraints", Fuzzy Sets and Systems, 91:165175, 1997.

[30] http://wenke.gtisc.gatech.edu/project/image004.gif

[31] Luo, J.,"Integrating fuzzy logic with data mining methods for intrusion detection", Master's thesis, Mississippi State Univ., 1999.

[32] Christos Douligeris, Aikaterini Mitrokotsa, "DDoS attacks and defense mechanisms: classification and state-of-the-art" ,Computer Networks: The International Journal of Computer and Telecommunications Networking, Vol. 44, Issue 5 , pp: 643 - 666, 2004.