Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2, Issue 1, January -2015

A Review on Classification Methods for Intrusion Detection System

Charli A. Solanki[1], Prof. Sheetal Mehta[2]

[1]Computer Science and Engineering, Parul Institute of Engg. And Technology, charli.solanki90@gmail.com
[2]Computer Science and Engineering, Parul Institute of Engg. And Technology, prof.sheetal.mehta@gmail.com

Abstract — *Security is becoming a critical part of organizational information systems. Intrusion Detection System (IDS) plays an effective role to achieve higher security in detecting malicious activities. The enormous volume of existing and newly appearing network data that require processing has given data mining classification and other techniques to make several important contributions to the field of Intrusion Detection. This paper includes various methods of classification which provide a path to detect intrusions with their advantages and disadvantages and comparative analysis.*

Keywords- *Intrusion Detection System, Classification, Data mining, Naïve bayes, Dataset.*

I. INTRODUCTION

Security is a big area of issue for all networks in today's environment. Hackers, attackers and intruders have made many successful attempts on company networks to break down there systems and web services. Many methods have been developed to secure the network infrastructure and communication over the Internet, among them the use of firewalls, encryption, and virtual private networks. Computer system is said to be reliable if confidentiality, integrity and availability is a part of its security requirements [1]. Intrusion detection is a relatively new addition to such techniques. Using intrusion detection methods, you can collect and use information from known types of attacks and find out if someone is trying to attack your network or particular hosts. The information collected this way can be used to harden your network security, as well as for legal purposes.

An intrusion is somebody attempting to break into or misuse your system. The word "misuse" is broad, and can reflect something severe as stealing confidential data to something minor such as misusing your email system for Spam [2]. An intrusion can be defined when any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource[3].

An **Intrusion Detection System** (IDS) is the device (or application) that monitors network/system activities and the analysing of data for potential vulnerabilities and attacks in progress, it also raises alarm or produces report [1]. Different sources of information and events based on information are gathered to decide whether intrusion has taken place. This information is gathered at various levels like system, host, application, etc. Based on analysis of this data, we can detect the intrusion based on two common practices – Misuse detection and Anomaly detection.

The main emphasis of this paper is to provide different classification methods for intrusion detection which provides advantages and disadvantages to choose appropriate approach. **In** this paper, different classification methods are there for detection of intrusions in network or host. Rests of the sections are as follows: Section 2 shows types of intrusion detection system. Section 3 evaluates different classification methods for IDS. Section 4 contains the comparison table with advantages and disadvantages and section 5 conclusions.

II. TYPES OF INTRUSION DETECTION SYSTEM

2.1 Detection Categories

2.1.1 Anomaly Detection

Normal system behavior is determined by observing the standard operation of the system or network traffic. An anomaly detection technique identifies the observed activities that deviate significantly from the normal usage as intrusions. As we shall see later, data mining techniques can be applied to determine if the system or network environment behavior is running normally or abnormally. Thus anomaly detection can detect unknown intrusions. The assumption in anomaly detection is that an intrusion can be detected by observing a deviation from the normal or expected behavior of the system or network [9].

2.1.2 Misuse Detection

Misuse detection IDS models function in very much the same sense as high-end computer anti-virus applications. That is, misuse detection IDS models analyze the system or network environment and compare the activity against signatures (or patterns) of known intrusive computer and network behavior [4]. This is also called signature

based detection. These signatures must be updated over time to include the latest attack patterns, much like computer anti-virus applications.

2.2 Primary Categories of IDS

2.2.1 Network based IDS

The network-based intrusion detection system (or NIDS) scans any traffic that is transmitted over the segment of the network and only permits through the packets that are not identified as intrusive.

2.2.2 Host based IDS

Host-based Intrusion Detection Systems are confined to monitoring activity on the local host computer. This monitoring can include network traffic to the host, or local object (files, processes, services) access on the host. For example, a HIDS implementation can be used to analyze all the network traffic transmitted to the computer and pass only the packets deemed safe onto the computer. A HIDS could also be a service running on the local machine that periodically examines the system security logs for suspicious activity..

The IDS performance is evaluated in term of accuracy, detection rate, and false alarm rate as in the following formula

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (1)$$

$$\text{Detection Rate (True Positive Rate)} = (TP) / (TP+FN) \quad (2)$$

$$\text{False Alarm} = (FP) / (FP+TN) \quad (3)$$

Where, TP (True Positive): Attack occurs and detected
False Positive (FP): Normal record predicted as attack
True Negative(TN): Normal record predicted as normal
False Negative(FN): Attack predicted normal

Table 2.1. Confusion Matrix for Evaluation

		Predicted Class	
		+(Normal)	**-(Intrusion /Attack)**
Actual Class	**+(Normal)**	TN	FP
	-(Intrusion/Attack)	FN	TP

Table 2.1 shows the categories of data behavior in intrusion detection for binary category classes (Normal and Attacks) in term of true negative, true positive, false positive and false negative.

III. CLASSIFICATION TECHNIQUES FOR INTRUSION DETECTION

In recent years, data mining techniques have been attracted by the researchers in the intrusion detection domain as they aim to reduce the great burden of analyzing huge volumes of audit data and producing optimization of detection rules [10].The term data mining is frequently used to designate the process of extracting useful information from large databases. There are a wide variety of data mining algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and databases. Intrusion detection can be thought of as a classification problem: we wish to classify each audit record into one of a discrete set of possible categories, normal or a particular kind of intrusions. We then use classification algorithm on audit data to build classifier. This classifier will then predict class of new unseen audit data as "normal" or "abnormal". Classification approach can be used for both misuse detection and anomaly detection but it is mostly used for misuse detection.

3.1 Decision tree

The decision tree model is used to classify data with common attributes. This algorithm consists of nodes, leaves, and edges. Every decision tree models starts with the root node and each node is an attribute data that decides which path will follow this node. An edge connects the nodes that attribute data assigns to. Moreover, this model makes the decision by comparing the value of data and labeled as leaves [6]. Its construction does not require any domain knowledge and it can also handle high dimensional data. Decision tree can process on both categorical and numerical data. It is understable easily get good accuracy but numeric dataset can be complex for algorithms.

3.2. Naïve Bayes [6]

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. It is fast in processing speed when applied to large databases. Naïve Bayes is one of the probabilistic approach for classification. It is widely used in knowledge management, image processing, and bio-

informatics etc. It calculates prior and conditional probabilities from a given dataset. Using these probabilities NBC find out the class value of unseen examples. The unseen example is then categorized to that class, which assumes the maximum probability.
Disadvantages are that results derived from threshold based systems are similar. Lack of available probability data is disadvantage for naïve bayes approach and it also depend on conditional independent attributes.

I.3. K- Nearest Neighbor [3]
K-Nearest Neighbor (k-NN) is a type of Lazy learning, it simply stores a given training tuple and waits until it is given a test tuple. For a given unknown tuple, a k-Nearest neighbor looks the pattern space for the k-training tuples that are closest to the unknown tuple. It is simplest among all machine learning algorithms. If k=1, then the object is simply assigned to its nearest neighbor. It is analytically traceable, Simple in implementation and uses local information and it lends itself easily to parallel implementation. With the disadvantages of large storage requirement. Highly susceptible to the curse of dimensionality and slow in classifying test tuples.

3.4. Neural Networks
A neural network consists of a collection of processing elements that are highly interconnected and transform a set of inputs to a set of desired outputs. The result is determined by the characteristics of the elements and the weights associated with the interconnections between them. By modifying the connections between the nodes, the network can adapt to the desired outputs. Neural networks have been used in both anomaly detection and misuse detection. For anomaly detection, neural networks were modelled to learn the typical characteristics of system users and identify significant variations from the user's established behaviour as anomaly. In misuse detection, the neural network would receive data from the network stream and analyze the information for instances of misuse [7]. It requires less formal statistical training. It can detect complex nonlinear relationship between dependent and independent variables or attributes. It can exhibit high tolerance to noise data. It's relatively greater computational burden and requires long training time.

3.5. Decision Table Majority Classifier
Decision Table is one of the possible simplest hypothesis spaces, and usually they are easy to understand. A decision table is an organizational or programming tool for the representation of discrete functions. It can be viewed as a matrix where the upper rows specify sets of conditions and the lower ones sets of actions to be taken when the corresponding conditions are satisfied; thus each column ,called a rule, describes a procedure of the type "if conditions, then actions". To build a decision table, the induction algorithm must decide which features to include in the schema and which instances to store in the body. So, some feature selection scheme is needed to be employed to select attribute that can be accommodated in schema for Decision Table Majority class ification [8].

IV. COMPARATIVE ANALYSIS OF VARIOUS CLASSIFIERS

Classifier	Parameters	Advantages	Disadvantages
Decision Tree	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.Decision tree Algorithms are unstable. 3.Trees created from numeric datasets can be complex.
Naïve Bayes	Class priors and feature probability distributions	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. 3. Give less accurate result.
Neural Network	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.	1. "Black box" nature. 2. Greater computational burden. 3. Requires long training time.

Classifier	Parameters	Advantages	Disadvantages
K-Nearest Neighbor	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. Slow in classifying test tuples.
Decision Table	Schema and Body containing instances with accurate attributes	1. Searches for exact matches using only the features in the schema 2. Highly accurate.	1.Takes time in processing. 2. Feature selection scheme is needed for attribute to accommodate in schema.

V. CONCLUSION

This survey describes various classification methods for intrusion detection using data mining techniques. Various classifiers have different way to solve the problems. As per the requirement and the parameters available this classifiers are used for effective intrusion detection system. So, our survey shows various classifiers that can be used with intrusion detection system. Moreover, advantages and disadvantages of various classifiers and the parameters that are used with the respective classifiers are listed.

REFERENCES

[1] Deepthy K Denatious, A John, "Survey on Data mining Techniques to Enhance Intrusion Detection", in International Conference on Computer Communication and Informatics, IEEE, pp. 10-12,2012.

[2] Allam Appa Rao, P.Srinivas, B. Chakravarthy, "A Java Based Network Intrusion Detection System (IDS)", Proceedings of The IJME – INTERTECH Conference, 2006

[3] Suraj S. Morkhade, Mahip Bartere, "Survey on Data Mining based Intrusion Detection Systems", in International Journal of Application or Innovation in Engineering & Management(IJAIEM), Vol. 2, Issue 3, March 2013.

[4] Wenke Lee, Salvotore J. Stolfo and Kui W. Mok, "A Data Mining Framework for Building Intrusion Detection Model", in Proceedings of the IEEE Symposium on Security and Privacy, pp. 120-132, 1999.

[5] Muhammad K. Asif, Talha A. Khan, Sufyan Yakoob, "Network Intrusion Detection and its Strategic Importance", IEEE 2013.

[6] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques ,Second Edition", ISBN 13: 978-1-55860-901-3, ISBN 10: 1-55860-901-6, 2006

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

[8] Ron Kohavi, "The power of decision Tables", in Eighth European Conference on Machine learning , pp. 174-189, 1995.

[9] E. Eskin, A. Arnold, M. Preau, L.Portnoy, and S. Stolfo, "A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data" in Applications of Data Mining in Computer Security, Kluwer Academic Publishers, 2002.

[10] P.Amudha, S.Karthik, S.Sivakumari " Classification Techniques for Intrusion Detection – An Overview" in International Journal of Computer Applications, Volume 76– No.16, August 2013.

