



SURVEY OF DATA MINING TECHNIQUES USED FOR INTRUSION DETECTION

B.SENTHILNAYAKI¹, M.CHANDRALEKHA², DR.K.VENKATALAKSHMI³
^{1,2}*Department of Information Technology, UCEVI*, ³*Department of Electronics And Communication Engineering, UCET*

ABSTRACT

With the rapid expansion of computer usage and computer network the security of the computer system has become very important. Every day new kind of attacks are being faced by industries. As the threat becomes a serious matter year by year, intrusion detection technologies are indispensable for network and computer security. A variety of intrusion detection approaches be present to resolve this severe issue but the main problem is performance. It is important to increase the detection rates and reduce false alarm rates in the area of intrusion detection. In order to detect the intrusion, various approaches have been developed and proposed over the last decade. In this paper, a detailed survey of intrusion detection based various techniques has been presented. Here, the techniques are classified as follows: i) papers related to Naïve bayes classifiers ii) papers related to Decision trees iii) papers related to KNN classifier iv) papers related to hybrid technique and v) paper related to other detection techniques. For comprehensive analysis, detection rate, time and false alarm rate from various research papers have been taken.

I.INTRODUCTION:

An Intrusion Detection System is a device or Software Application that monitor the networks or System Activities for malicious activities or policy violation and produces report to a management station. The Intrusion Detection System comes in variety of flavors and approach the goal of detecting suspicious traffic in the different way. There are network based (NIDS) and host based (HIDS) Intrusion detection system.

Some of the systems may attempt to stop an Intrusion but this is neither require nor expected by a Monitoring Systems. The Intrusion Detection Prevention System are primarily information about them and reporting the attempts. The many Organizations use this for purposes such as problems with the Security policies, documenting existing threats. Network Intrusion Detection System (NIDS) are placed at a point or points within the network to monitor the traffic to and from all devices. Once an attack is identified or any abnormal behavior is sensed, the alert is send to the administrator.

The attack may be a passive or also an active attack.

Statistical anomaly- Based IDS monitor the network traffic and compare with its baseline. The Baseline defines what network, Bandwidth, protocols and what ports an devices should be connected, How and when the traffic should be monitored. An IDS is categorized as behavior-based system, when it uses information about the normal behavior of the system it monitors. Behavior on detection describes the response of the IDS after the detection of attacks. It can be divided into active or passive based on the attack response. These two types of intrusion detection systems differ significantly from each other, but complements one another well. The architecture of host-based is completely dependent on agent- based, which means that a software agent resides on each of the hosts, and will be governed by the main system.

Intrusion detection products are tools to assist in managing threats and vulnerabilities in this changing environment. Threats are people or groups who have the potential to compromise your computer system. These may be a curious nemesis to many companies in the software field More efficient host-based intrusion detection systems are capable of monitoring and

collecting system audit trails in real time as well as on a scheduled basis, thus distributing both CPU utilization and network overhead and providing for a flexible means of security administration. It would be advantageous in IDS implementation to completely integrate the NIDS, such that it would filter alerts in a identical manner to HIDS and can be controlled from the same environment should require both NIDS and HIDS to be implemented for not only providing a complete defense against dynamic attacks but also to effectively and computer/network misuse against threats and malicious activities.

Users choose or are assigned an ID and password or other authenticating information that allows them access to information and programs within their authority. Network security covers a variety of computer networks, both public and private, that are used in everyday jobs conducting government agencies and individuals. Networks can be private, such as within a company, and others which might be open to public access. Network security is involved in organizations, enterprises, and other types of institutions. It does as its title explains: It secures the network, as well as protecting and overseeing operations being done.

The most common and simple way of protecting a network resource is by assigning it a unique name and a corresponding password. Network security starts with authenticating, commonly with a username and a password. Since this requires just one detail authenticating the user name i.e. the password this is sometimes termed one-factor authentication. With two factor authentication, something the user 'has' is also used (e.g. a security token or 'dongle', an ATM card, or a mobile phone); and with three-factor authentication, something the user 'is' is also used (e.g. a finger print or retinal scan). Once authenticated, a firewall enforces access policies such as what services are allowed to be accessed by the network users.

Though effective to prevent unauthorized access, this component may fail to check potentially harmful content such as computer worms or Trojans being transmitted over the network. Anti-Virus software or an Intrusion Prevention System (IPS) help detect and inhibit the action of such malware. An anomaly based intrusion detection system may also monitor the network like wireshark traffic and may be logged for network, as well as protecting and overseeing operations being done. Most definitions of network security are narrowed to the enforcement mechanism. Enforcement concerns analyzing all network traffic flows and should aim to preserve the confidentiality, integrity, and availability of all systems and information on the network. These three principles compose the CIA triad:

- Confidentiality - involves the protection of assets from unauthorized entities
- Integrity - ensuring the modification of assets is handled in a specified and authorized manner
- Availability - a state of the system in which authorized users have continuous access to said assets.

DATA CLASSIFICATION ALGORITHMS:

Classification is a model finding process that is used for portioning the data into different classes according to some constraints. In other words we can say that classification is process of generalizing the data according to different instances. Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs.

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. For example, you may want to predict whether individuals can be classified as likely to respond to a direct mail solicitation, vulnerable to switching over to a competing long-distance phone service, or a good candidate for a surgical procedure. People are often do mistakes while analyzing or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to successfully applied to these problems, improving the efficiency of systems and the designs of machines. There are several applications for Machine Learning (ML), the most significant of which is data mining. Numerous ML applications involve tasks that can be set up as supervised. In the present paper, we have concentrated on the techniques necessary to do this. In particular, this work is concerned with classification problems in which the output of instances admits only discrete, unordered values.

Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long-distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. For example, a sample of a mailing list would be sent an offer, and the results of the mailing used to develop a classification model to be applied to the entire

database. Sometimes an expert classifies a sample of the database, and this classification is then used to create the model which will be applied to the entire database.

NAÏVE BAYES CLASSIFIERS:

A many papers have been presented to represent the native bayesian classifier based intrusion detection. Some of the papers have been discussed below. Text classification and the feature selection is used to improve accuracy for the text classifiers in the preprocessing technology. Domain and algorithm are considered in the feature selection. The Bayesian algorithm is very simple and very efficient.

Houkuan Huang et al stated that Naïve bayes is highly sensitive to feature selection. So that the research on this I very important. Two evaluation metrics are carried on multi- class text datasets: Multi-class Odds Ratio (MOR), and Class Discriminating Measure (CDM) [1]. The experiments on two multiclass tests collection were carried on. The results is the better feature selection than other approaches.

Naive Bayes is a widely used classification method based on Bayes theory. Based on class conditional density estimation and class prior probability, the posterior class probability of a test data point can be derived and the test data will be assigned to the class with the maximum posterior class. The key problem in naive Bayes method is the class conditional density estimation. A Latest approach to the alert classification to reduce false positives is using improved self adaptive Bayesian algorithm (ISABA) [3]. The class conditional density is estimated based on data points. For uncertain classification problems, we should learn the class conditional density from uncertain data objects represented by probability distributions.

Jiangtao Ren et al stated that the naïve Bayes method is extended to handle uncertain data, three methods are used [2]. Averaging (AVG): We first obtain the average point of every uncertain data object. Then, these points are passed to naive Bayes. Sample-based method (SBC): The kernel function, which is the key function used in naive Bayes, is redesigned to consider values sampled from the uncertain data as input. In this method, the probability distributions can be arbitrary. Formula-based method (FBC): This is a special application probability. of the sample-based method, where a closed-formula for the kernel function is derived. We have derived the formula for Gaussian distribution. (Lacking space, we omit the results for uniform distribution.)

A novel naive Bayes classification algorithm for uncertain data with a pdf (probability distribution function). Our

key solution is to extend the class conditional probability estimation in the Bayes model to handle pdf's. Extensive experiments on UCI datasets show that the accuracy of naïve Bayes model can be improved by taking into account the uncertainty information.

K NEAREST NEIGHBOUR CLASSIFIERS:

A Many papers have been presented to represent the k Nearest neighbour classifier based intrusion detection. Some of the papers have been discussed below. The difficult in the probability is sampling and deciding the membership of pattern to be classified and its typicalness. The fuzzy is introduced in the K Nearest neighbour algorithm to develop the algorithm [4]. Three methods of sampling are proposed. It resulted in low error rate and confidence measure of classification. It is very efficient than other standard, more sophisticated pattern recognition procedures in these experiments.

K Nearest Neighbor is one of the proposed method. The performance of the kNN on different data sets gave a good results. The disadvantages of the traditional kNN are: (i), calculation complexity due to the usage of all the training samples for classification, (ii) the performance is solely dependent on the training set, and (iii) there is no weight difference between samples. To solve this disadvantage of kNN it is proposed with Genetic Algorithm [5]. Instead of considering all the training samples and taking k-neighbors, the GA is employed to take k-neighbors straightaway and then calculate the distance to classify the test samples. The reduced data set received from Rough set theory with Bee Colony (BCO) is classified. The performance is good than the CART and SVM classifiers.

Though kNN (k Nearest Neighbor) is effective in result in larger model sizes. Many Algorithms are used to reduce the size of kNN for the purpose of accuracy [6]. Toh Koon Charlie Neo et al stated that a direct boosting algorithm for the k-NN classifier that creates an ensemble of models with locally modified distance weighting. A study is conducted on ten standard databases from UCI respository show that the kNN for the accuracy of the datasets. Diego P. Vivencio et al stated that a feature weighting method based on χ^2 statistical test, to be used with k-NN classifier [7]. The Results of Experiments with various data from different domains are discussed. It proposed that it is a good weighting strategy.

DECISION TREE CLASSIFIERS:

A Many papers have been presented to represent the Decision tree classifier based intrusion detection. Some of the papers have been discussed below. The learning algorithm faced many problems in selecting the subsets

which to focus its attention which ignore the rest. Ron Kohavi et al states that to improve performance we should consider how algorithm and training data sets should interact [8]. The relation between the feature subset selection and relevance is explored. George et al states that the data with values arises the value scepticism in numerous applications [9]. The study of strength and weakness is discussed. The filter approach to feature subset selection and wrapper approach is compared and studied. Improved accuracy is achieved for some of the data sets.

OTHER CLASSIFIERS:

An approach to network interference detection, based merely on a hierarchy of Self-Organizing Feature Maps has been inspected by H. Gunes Kayacik *et al.* [10]. Establishing immediately how far such an approach was full in practice was their principle interest. For doing this, the KDD benchmark dataset from the International Knowledge Discovery and Data Mining Tools Competition was utilized. Extensive analysis was conducted In order to concentrate on the importance of the features used, the division of training data and the complication of the architecture. Using unsubstantiated learning in comparison to results statement formerly. They explained that most excellent presentation was obtained by means of a two-layer SOM hierarchy, based on all 41-features since the KDD dataset.

Yang Li and Li Guo have presented a paper Based on TCM- KNN (Transductive Confidence Machines for K-Nearest Neighbors) machine learning algorithm and dynamic education based training data selection scheme a supervised network intrusion detection method in the year 2007 [11]. It was successfully identified anomalies with elevated detection rate, low false positives under the condition of utilizing much less chosen data as well as selected features for training in association with the traditional managed intrusion detection techniques. The recommended method was more tough and successful than the state-of-the-art intrusion detection methods which were explained by a chain of experimental results on the familiar KDD Cup 1999 data set.

In advance to the above concept in the year 2009, a network based anomaly detection system that makes use of a hierarchy of SOMs has been offered by Saroj Kumar Panigrahy *et al.* [12]. By means of a controllable rate of false alarms the system was set up to identify very soon over 60% of the attacks. Even though the result of this job was construed with warning, it was recommended that the arrangement offered carry out comparably to few of the superior systems that took part in the DARPA Intrusion Detection Evaluation. The system was also not at all trained on the complete training dataset, sense that it

could not have had a opening to learn the full sort of usual behavior and it was not tested on the full test dataset, i.e., it may not have come across few of the more not easy attacks.

In development of the above concept in the year 2010 by means of enhanced self adaptive Bayesian algorithm (ISABA), an approach to the vigilant classification to lessen false positives in intrusion detection has been offered by Dewan Md. Farid and Mohammad Zahidur Rahman [13]. The recommended approach used to the security domain of anomaly based network intrusion detection, which properly categorized dissimilar types of attacks of KDD99 benchmark dataset with elevated arrangement rates in small reply time and decrease false positives with restricted computational assets. In 2010, an intrusion detection system was constructed by Muna M.

Taher Jawhar and Monica Mehrotra [14] using hamming and MAXNET Neural Network for recognizing attack class in the network traffic. One more approach based on Multilayer Perceptrons (MLP) network has been derived and the evaluation of the system is done by comparing the results of the two approaches. The results of experiments confirmed that the designed models were capable in terms of accuracy and computational time of real word intrusion detection. Defense Advanced Research Projects Agency (DARPA) intrusion detection evaluation datasets provided the necessary training and testing data. The utilization of Self Organizing Maps for building an Intrusion Detection System is well described by Mr. Vivek A. Patole *et al.* [15] in the same year. The system architecture and the flow diagram for the SOM have been explained.

The advantages and disadvantages of the algorithm have also been presented. Their real experiments proved that even a simple map, when trained on usual data, could detect the anomalous features of both buffer overflow intrusions to which they are exposed. Also, an interference detection system with Bayesian probability has been developed Saeed Algarny et al [16]. To categorize possible intrusions, the structure developed was an adolescent Bayesian classifier that was utilized. The structure was taught a priori by means of a subset of the KDD dataset. The capability of Bayesian classifier was to distinguish the interference with a better detection rate.

With Conditional Random Fields and Layered Approach, Kapil Kumar Gupta et al. [17] had lectured two topics of Accuracy and Efficiency. They revealed high attack detection precision was obtained by means of Conditional Random Fields and high competence by applying the Layered Approach. Their recommended system based on Layered Conditional Random Fields outperforms other familiar means such as the decision trees and the naive

Bayes was shown by the experimental results on the benchmark KDD '99 intrusion data set. Detection accuracy for our method is confirmed by statistical Tests which induced higher confidence level. In conclusion, it has been shown that the system was healthy and is capable of handling noisy data with no compromise on performance

In development to the above approach, for independent rule creation, a multi-modal genetic algorithm solution has been discovered by Todd Vollmer et al. [18]. Relatively than the development of rules to give a solution for interference detection this algorithm spotlight on the process of making rules once interference had been recognized. The algorithm was explained on irregular ICMP network packets (input) and Snort rules (output of the algorithm). According to a fitness value output rules were arranged and any replacements were detached.

These rules were precise to the packets and formed only four false positives from 33,804 test packets that were shown by testing. In advance to the above technique in 2011, different ways such as to categorize normal and intrusive actions or to extract interesting intrusion models, Machine learning techniques are often used to interference detection problems. Besides providing interference classification capacities self learning rule based systems can ease area specialists from the complicated task of hand crafting signatures.

A genetic-based signature knowledge system that was adaptively and energetically learns signatures of both standard and interfering actions from the network traffic has been developed by Kamran Shafi, Hussein A. Abbass et al [19] to this end. The assessment of their systems to actual time network traffic which was incarcerated from a university departmental server was completed by them. By combining real background traffic with attacks replicated in a controlled environment an attitude was developed to put together fully labeled interference detection data set. Proper for a managed learning classifier system and other associated machine learning systems apparatus were improved to preprocess the unrefined network data into quality vector format. The signature extraction system was then pertained to this data set.

CONCLUSION:

Network Intrusion Detection System is a latest kind of defense technology which is one of the vibrant areas in network security. In recent years many techniques are available for intrusion detection. In this paper, a detailed survey of important techniques based on intrusion detection is presented. Also the classification of the techniques based on neural network, k-means, hybrid

techniques, support vector machine etc., is provided. Detection rate and false alarm rate are considered for comprehensive analysis

\

REFERENCES:

1. Jingnian Chen "Feature selection for text classification with Naïve Bayes", *Expert Systems with Applications* vol.36, 2009, pp.5432–5435.
2. Jiangtao Ren "Naive Bayes Classification of Uncertain Data", *IEEE International Conference on Data Mining*, 2009
3. Dewan Md. Farid "Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm", *Journal Of Computers*, vol. 5, NO. 1, January 2010, pp.23- 31.
4. James M.Keller, "A Fuzzy k-Nearest Neighbor Algorithm", *IEEE Transaction on Cybernetics* vol.15, August 1985, pp.580-585.
5. N. Suguna " An improved k – Nearest Neighbore Classifications Using Genetic algorithm", *International Journal of Computer Science Issues*, Vol. 7, Is sue 4, No 2, July 2010, pp.8-21.
6. Toh Koon Charlie Neo "A direct boosting algorithm for thek-nearest neighbor classifier via local warping", *Pattern Recognition Letters*, vol.33, 2012, pp. 92–102.
7. Diego P. Vivencio "Feature-weighted k-Nearest Neighbor Classifier", *IEEE Symposium on Foundations of Computational Intelligence*, 2007.
8. Ron Kohavi "Wrappers for feature subset selection", *Artificial Intelligence*, vol. 97, 1997, pp.273-324
9. Kiran Siripuri "Classification of Uncertain Data using Decision Trees", *International Journal of Advanced Research In Computer Science and Software Engineering*, vol.3, 2013, pp. 231-239.
10. H. Gunes Kayacik, A. Nur Zincir-Heywood, Malcolm Heywood, "A Hierarchical SOM based Intrusio Detection System, "Engineering Applications of Artificial Intelligence, Vol.20, no.4, pp.439-451, June 2007.
11. Yang Li, Li Guo, "An active learning based TCM-KNN Algorithm detection," *computers & security*, vol.26, pp.459-467, 2007.
12. Saroj Kumar Panigrahy, Jyoti Ranjan Mahapatra, Jignyanshu Mohanty and Sanjay Kumar Jena, "Anomaly Detection in Ethernet Networks using Self Organizing Maps," *Department of Computer Science*, 2009.
13. Dewan Md. Farid, Mohammad Zahidur Rahman, "Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm," *Journal of Computers*, Vol. 5, No. 1, January 2010.
14. Muna M. Taher Jawhar and Monica Mehrotra, "Anomaly Intrusion Detection System using Hamming Network Approach," *International Journal of Computer*

Science & Communication, Vol. 1, No. 1, pp. 165-169, January-June 2010.

15. Mr. Vivek A. Patole, Mr. V. K. Pachghare, Dr. Parag Kulkarni, "Self Organizing Maps to Build Intrusion Detection System," *International Journal of Computer Applications*, pp. 0975-8887, Vol. 1, No. 8, 2010.

16. Hesham Altwaijry, Saeed Algarny, "Bayesian based intrusion detection system," *Journal of King Saud University, Computer and Information Sciences*, 2010.

17. Kapil Kumar Gupta, Baikunth Nath, and Ramamohanarao Kotagiri, "Layered Approach Using Conditional Random Fields for Intrusion Detection," *IEEE Transactions On Dependable And Secure Computing*, Vol. 7, No. 1, January-March 2010.

18. Todd Vollmer, Jim Alves-Foss, Milos Manic, "Autonomous Rule Creation for Intrusion Detection," *IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*, pp. 1-8, 2011.

19. Kamran Shafi, Hussein A. Abbass, "Evaluation of an Adaptive Genetic-Based Signature Extraction System for