A Comprehensive Survey of Selected Data Mining Algorithms used for Intrusion Detection

Vivek Kumar Srivastava
MTech(CSE), CDAC- Noida
GGSIP University
Delhi, India
viveks7070@gmail.com

Abstract: Internet is widely spread in each corner of the world, so there is always a possibility of unauthorized attacks. To protect the computers from these unauthorized attacks, effective intrusion detection systems (IDS) need to be employed. In recent years, various data mining approaches have been used with Intrusion Detection System to improve the accuracy and detection of novel types of intrusion. This paper evaluates the performance of various well known methods to classify the normal and attacked data, and finally advantage and disadvantage of various data mining algorithms are discussed. It has been observed that many challenges still exist in the design and implementation of IDS.

Key Words: Data Mining, Decision tree, Fuzzy logic, Genetic algorithm, Intrusion Detection K-mean, Knowledge discovery database (KDD CUP99 data set), Support Vector Machine.

1. INTRODUCTION

An Intrusion Detection System (IDS) was first introduced in 1980 by James Anderson. A sufficient amount of research has been done on intrusion detection technology but still it is considered as immature and not perfect against intrusion. Without using data mining, IDS requires a significant amount of human work. Data mining based IDS takes less expert knowledge yet provides good performance with known and unknown attacks [10]. IDS is placed centrally to capture all incoming packets that are transmitted over the network. Data are collected and sent for pre-processing to remove the noise; irrelevant and missing attributes are replaced. Then the preprocessed data are analyzed and classified according to their severity measures. If the record is normal, then it does not require any more change and if records have anomaly, it is sent for report generation. Based on the state of the data, alarms are raised and information goes to the administrator to handle the situation in advance [2].

1.1 Intrusion Detection

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect malicious activities taking place through the network. ID is an area growing in significance as more and more sensitive data are stored and processed in networked systems. Each malicious activity or attack has a specific pattern, the patterns of only some of the attacks are known whereas the other attacks only show some deviation from the normal patterns [10] [11]. Therefore, the techniques used for detecting intrusions are based on whether the patterns of the attacks are known or unknown. The two main techniques are used: misuse detection and anomaly detection. Misuse detection is based on the knowledge of known patterns of previous attacks and system vulnerabilities. Misuse detection continuously compares current activity to known intrusion patterns to ensure that any attacker is not attempting to exploit known vulnerabilities. To accomplish this task, it is required to describe each intrusion pattern in detail. It cannot detect unknown attacks. Anomaly detection, on other hand is based on the assumption that intrusion always reflects some deviation from normal pattern. The normal state of the network, traffic load breakdown, protocol and packet size are defined by the system administration in advance, and thus anomaly detector compares the current state of the network to the normal behavior and looks for malicious behavior. It can detect known and unknown

attacks [2] [3]. Generally, there are four categories of attacks [2].

D o S – Denial of Service: In this type of attacks intruder makes processing time of the resources and memory busy so as to avoid true user from accessing those resources.

Probe – Surveillance and probing: Intruder examines a network to find out well-known vulnerabilities of the target machine. These network analysis are reasonably valuable for an attacker who is planning an attack in future. For example: port-scan etc.

R2L – Remote to Local: Unauthorized attackers get local access of the target machine from a remote location and then feat the target machine vulnerabilities. For example: guessing password etc.

U2R – User to Root: Target machine is already attacked, but the intruder try to gain access with super-user privileges. For example: buffer overflow attacks etc.

1.2 *Data Mining*

In IDS, data mining plays a vital role to determine intruded and normal data. Data mining framework automatically detect patterns in our data set and use these patterns to find a set of malicious activity. By applying data mining technology, intrusion detection system can widely verify the data to obtain a model, thus helps to obtain a comparison between the abnormal pattern and the normal behavior pattern. Manual analysis is not required for this method. One of the main advantages is that same data mining tool can be applied to different data sources [10][11].

1.3 **KDD CUP99 Data se**

The KDD CUP99 dataset is a benchmark dataset which was simulated in military network environment in 1998 then derived to KDD99 dataset in 1999. Knowledge discovery in databases (KDD) is used to denote the process of extracting useful knowledge from large data sets. The main reason behind for selecting KDD dataset is that currently, it is the most widely used comprehensive data set that is shared by many researchers. In this dataset, 41 attributes are used in each record to characterize network traffic behavior. Among these 41 attributes (table 1), 3 are nominal and 38 are numeric attributes. All features in KDD dataset are grouped into three categories a traffic features, Content features, Basic features [9] [10].

Table 1: List of features available in KDD CUP 99 dataset

S.No	Feature Name	S.No	Feature Name
1	Duration	22	is_host_login
2	protocol_type	23	is_guest_login
3	Service	24	Count
4	Flag	25	srv_count
5	src_bytes	26	serror_rate
6	dst_bytes	27	srv_serror_rate
7	Land	28	rerror_rate
8	wrong_fragment	29	srv_rerror_rate
9	Urgent	30	same_srv_rate
10	Hot	31	diff_srv_rate
11	num_failed_logins	32	srv_diff_host_rate
12	logged_in	33	dst_host_counte
13	num_compromised	34	dst_host_srv_count
14	root_shell	35	dst_host_same_srv_rate
15	su_attempted	36	dst_host_diff_srv_rate
16	num_root	37	dst_host_same_src_port_rate
17	num_file_creations	38	dst_host_srv_diff_host_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_files	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_rerror_rate
21	dst_host_serror_rate		

2. SURVEY OF VARIOUS DATA MINING TECHNIQUES FOR IMPLEMENTING INTRUSION DETECTION SYSTEM

This section presents a survey of various data mining techniques that have been applied to IDS by various research groups. Data mining techniques can be differentiated by their different model functions and representation, preference and algorithms. This paper is mainly focused at classification of normal and malicious data. Common representations for data mining techniques include rules, decision trees, feature selection, genetic algorithms, fuzzy logic, support vector machine, neural network, clustering techniques.

2.1 Decision Tree

Decision tree is predictive modeling technique mostly used for classification in data mining. It is a form of data analysis which takes each instance of a dataset and assigns it to a particular class. It is two step processes, in the first step; a classifier is build describing a predefined set of data classes or concepts. This is a learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning form" a training set made of database tuples and their associated class labels. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached [2] [14]. Using decision tree algorithms for IDS is that its construction does not require any domain knowledge. Hence a data mining expert with little knowledge of networking can help build accurate decision tree models. Another significant advantage is that decision trees can handle high dimensional data [14] [15]. This increases the suitability of decision tree algorithms for IDS especially while considering the heterogeneity of network connection data and its ever increasing size. Decision trees are able to process both numerical and categorical data.

G.V. Nadiammai, M. Hemalatha [2] worked on the problem of classification of data by introducing data adopted decision tree algorithm. This proposed decision tree is different from normal decision tree algorithm. It efficiently classifies the data into normal and attack. To optimize the result of decision tree, author used PSO(Particle Swarm Optimization) technique that identifies global and local best value for n number of iterations to obtain the optimal solution. The best solution is obtained by calculating the average value and by finding the exact efficient features from the given training data set. The proposed EDADT(Efficient Data Adopted Decision Tree) algorithm drastically improve the detection rate by 98.12% and is also very effective to improve the false alarm rate by 0.18%. All of these experiments are perform on KDD CUP 99 data set.

Vaishali Kosamkar, Sangita S Chaudhari [3] proposed a hybrid algorithm to develop the IDS. In this hybrid approach, author used C4.5 decision tree with SVM (Support Vector Machine) to achieve high detection rate and diminish false alarm rate. Experimental result shows that there is increase in accuracy and detection rate whereas false alarm rate is reduced. In this hybrid approach, author takes advantage of both C 4.5 and SVM algorithm in such a way that C 4.5 gives high detection rate and SVM gives low false alarm rate. The proposed hybrid algorithm is able to increase the accuracy of IDS with 98.30% and simultaneously decrease the false alarm rate with 1.01%. It is concluded that using SVM with C 4.5 improves the working of an IDS.

2.2 Fuzzy Logic

Fuzzy logic (Zadeh 1965) works with reasoning rules very close to the human way of thinking, which is approximate and intuitive. The main characteristic of fuzzy logic is that it allows us to define values without specifying a precise value, something which is not possible with classical logic, upon which computer development has been based so far [12] [7].

P. Jongsuebsuk. Wattanapongsakorn, C. Charnsripinyo [1] proposed fuzzy logic with genetic algorithm to optimally classify the network data. In this paper testing attributes are normalized between the real number 0.0 to 7.0, where the maximum and minimum values of testing data set are 7.0 and 0.0 respectively. The authors found the probability of each record for each detection rule and count for true positive (Number of attack record which are correctly classified as attacks) and true negative (Number of normal record which are correctly classified as normal records). Trapezoidal shape is used to measure the probability of being attack identified by each attribute. Genetic algorithm is used to optimize the result produced

by fuzzy rules. After counting all records from all rules, fitness value for each detection rule is calculated and rule with highest fitness value is preserved. They performed their experiment on 26,500 records of online network traffic and result shows that they got 98.72% detection rate and low false alarm rate. Previously, detection rate was around 93%, thus their approach worked well to enhance detection rate.

Roya Ensafi, Soheila Dehghanzadeh, Mohammad -R. Akbarzadeh [7] proposed a hybrid and novel method to efficiently identify the normal and intrudes data. They proposed SFK-means approach which inherits advantage of k-means, fuzzy k-means and swarm optimization. This paper well solved the local convergence problem by fuzzy k-means and sharp boundary problem by swarm k-means. Proposed method consists two phase, training and testing phase. In training phase, classification of data is done which is impressed by best particle parameter through the generations. In detection phase, Euclidian distance between cluster centroids of the evolved particle and the input data is calculated to find out if the input data is normal or anomaly. After applying SFK-means algorithm they achieved 95.876% overall detection rate and 2.1247% false alarm rate. They also concluded from the experiment that 98.9% detection rate and 2.147% false alarm rate is achieved for DoS type of attack, 87.76% detection rate and 0.938% FAR for R2L type of attacks, 98.432% detection rate and 0.939% FAR for U2R type of attacks and 95.43% detection rate and 1.70% FAR for Probes types of attacks. Analysis of result shows that their proposed approach worked well for U2R and R2L type of attacks.

2.3 Genetic optimization

Genetic algorithm is a search methodology that copies the process of natural selection. This method is iteratively generating useful solution to optimization. Genetic belong to class of evolutionary algorithm, which generate solutions to optimization problem inspired by natural evolution such as selection, crossover and mutation [9].

B.Senthilnayaki, K.Venkatalakshmi, A. Kann an [9] proposed an intelligent algorithm for intrusion detection. The system uses a new genetic algorithm for preprocessing and modified J48 algorithm is used to identify intended activities. The noticeable contribution of this paper is that they proposed a new processing technique for removing noisy data in KDD CUP 99 dataset. This paper consists of eight major modules namely dataset, pre-processing module, optimal feature selection, classification sub system, decision manager, rule manager, administrator and knowledge base. Genetic algorithm is used for optimal feature selection with the help of mutation and crossover operation. Using genetic algorithm, the nine features (protocol type, service, src_bytes, dst_bytes. Flag, diff_srv_rate, dst_host_serve_count, dst_host_serror_rate, dst_host_srv_serror_rate) out of 41 features of KDD CUP 99 dataset have been selected. Modified J48 classification algorithm is used to classify normal and attacked data, and from the experimental result it has been observed that the training and testing time are reduced in the proposed modified J48 algorithm. After applying modified J48 algorithm on KDD CUP 99 data set around 98% detection rate and 1.5% false alarm rate is achieved. The best part of this approach is that the decision tree is not applied on complete data set so time is reduced to generate the decision tree and make conclusion for normal and intrudes data.

2.4 Support Vector Machine

Support Vector Machines are the supervised learning methods used for regression and classification, SVM belong to class of linear classifiers. SVMs separate the data into multiple classes though the use hyper-plane. An advantage of SVM as a classifier for an IDS is that they are highly accurate, it has one more advantage it easily able to model complex nonlinear decision boundaries and are less prone to over fitting than other method. A disadvantage of SVM as a classifier is high degree of complexity and large memory requirements due to this speed of both testing and training are slow [3].

Vaishali Kosamkar, Sangita S Chaudhari [3] proposed a hybrid method for an IDS in which they combined decision tree and SVM to get high detection rate and low false alarm rate. The result of C 4.5 decision tree gives input to SVM where it tries to solve the binaries classification problem. Later it maps linear algorithm to nonlinear space.

2.5 Clustering Technique

Data clustering is an unsupervised learning model and is a common technique for statistical data analysis, which is used in many fields of data mining, pattern recognition and biometrics. Clustering is partitioning of data set into subset so that data in each subset hold similar property. It has great advantage over classification that it does not require prior knowledge of dataset for training [5] [14].

LI Han [5] proposed a method which tries to sort out some problems which are present in k-means algorithm, like choosing initial seed problem, presence of outlier and noise. Author has made a very nice attempt to sort out these problems. To deal with outlier and noise, for each point, they calculated sum of distance between that point and all other points and compared it with sum of average distance. To find out appropriate initial centroid, concept of density is used. From the experimental results it is clear that proposed approach enhanced the working of an IDS. Before applying improvement in K-mean, detection rate was around 95% and FAR was around 4%, however, after doing improvement in K-means, result is improved and detection rate become 96.5% and false alarm rate goes to 2.1%. Thus proposed approach has enhanced the working of IDS effectively.

3. RESULT AND DISCUSSION

A summarized result of well known six algorithms are presented in table 2, from the result we conclude that decision tree (J48; java implementation of C 4.5) gave the best detection rate, but it has disadvantage that it will crisply classifies the incoming packets which lead sometimes false alarm rate, for this fuzzy rule sets are used to sort out the problem of decision tree, it has another drawback of local optimal problem to get the advantage of global optimization; genetic and particle swarm optimization algorithm are introduced with decision tree.

3.1 Measurement Terms

Detection rate and False Alarm Rate have been used for measure the performance of various techniques that are implemented to IDS.

Detection Rate (DR): Percentage of attacked data which are correctly classified as an attack data.

DR= (TP/ (TP+FN)) * 100

False Alarm Rate (FAR): Percentage of normal data which are incorrectly classified as an attack data.

FAR= (FP/ (FP+TN)) * 100

Where TP (true positive) is number of attack data which are correctly classified as attack, TN (true negative) is number of normal records which are correctly classified as normal, FP (false positive) is number of normal records which are incorrectly classified as attack, FN (false negative) is number of attack records which are incorrectly classified as normal.

4. CONCLUSION AND FUTURE WORK

This paper presents a survey on various data mining algorithms that have been used for improvement in Intrusion Detection System. The results have been evaluated by researchers and it is important to see that each of the approaches has its own pros and cons; Table 2 shows the detection and false alarm rate of these algorithms. It has been observed that instead of using single algorithm, combination of algorithms have been used to get high detection rate and low false alarm rate [1] [2] [3] [7]. Future work in this context involves exploring new hybrid methods for enhancing the performance of IDS. Till now, work has been done mainly focusing offline traffic, so in future online traffic can also be focused.

Table 2: Comparison of various data mining algorithm

Algorithm	Approach	Detection Rate	False Alarm Rate
J48 Classification algorithm	Modified J48 algorithm with genetic algorithm	98.80%	1.05%
Decision tree and SVM	Hybrid algorithm based on C 4.5 decision tree and Support Vector Machine	98.623%	1.01%
K-means	A new clustering method based on K-means algorithm	Neptune attack: 96.10% Buffer_overflow attack: 98% Guess_password attack: 99% Portsweep attack: 97.50%	1.9% 3.9% 0.6% 2.2%
Fuzzy Genetic Algorithm	Fuzzy rule set with Genetic Optimization	98.72%	FP:0.13% FN:1.55%
Fuzzy K-means	Fuzzy K-means with Particle Swarm Optimization	95.876%	2.1247%
Decision Tree	Decision tree with Particle Swarm Optimization	98.12%	0.18%

REFERENCES

[1] *P.jongsuebsuk, N.Wattanapongdakorn, and C.Charnsripinyo, "Improved IDS with Fuzzy Genetic Algorithm", IEEE UMM Csci Senior Seminar conference, December2013.*

[2] *G.V.Nadiammai and M.Hemalatha, "Effective approach towards IDS using Data Mining", Egyptian Informatics journal production and hosted by Elsevier- 2013.*

[3] *Vaishali Kosamkar and Sangita Schaudhari, "Improved IDS using C 4.5 decision tree and SVM", International journal of computer science and information technology -2013.*

[4] *Idown S.A, Anyaechie Amarachi, Ajayi, "Compararive Study of Selected Data mining algorithm used for IDS", IJSCE 2013.*

[5] *B.Senthilnayaki, K.Venkatalakshmi, A. Kann an, "An Intelligent Intrusion Detection System using Genetic Based Feature Selection and Modified J48 Decision Tree Classifier", IEEE fifth international conference on Advanced Computing (ICoAC), December 2013.*

[6] *Ajayi Adebowale, Idowu S.A, Anyaehie Amarachi, "Comparative Study of Selected Data Mining Algorithms Used For Intrusion Detection", International Journal of Soft Computing and E engineering (IJSCE), July 2013.*

[7] *Chih-Fong Tsai,Jung-Hsiang Tsai,Jui-Sheng Chou, "Centroid –Based Nearest Neighbor Feature representation for Intrusion detection", The Institute of Electronics, Information and Communication Engineeres, IEICE- 2012.*

[8] *E.Kesavulu Reddy, Member IAENG, V.Naveen Reddy, P.Govinda Rajulu, "A Study of Intrusion Detection in Data Mining", World Congress on Engineering IEEE. July 6 - 8, 2011, London, U.K.*

[9] *LI Han, "Research and Implementation of an Anomaly Detection Model Based on Clustering Analysis", International Symposium on Intelligence Information processing and trusted computing IEEE-2010.*

[10] *T.P., Fries, "Evolutionary optimization of a Fuzzy rule Based network Intrusion Detection system", IEEE Conference on Fuzzy Information Processing Toronto, July 2010.*

[11] *Roya Ensafi,Soheila Dehghanazadeh, Mohammad R. Akbarzadeh, "Optimizing Fuzzy K-means for Network anomaly detection using PSO", International Conference on Hybrid information Technology, IEEE, 2008.*

[12] *Te-Shun and Kang K. Yen, "Fuzzy Belief K-nearest neighbor anomaly detection", Workshop on Information Assurance United States Military academy, IEEE, West point ,Ny ,June 2007.*

[13] *KDD cup99 Intrusion detection dataset.http://kdd.ics.uci.edu/databases/kddcup99.html.*

[14] *"Data Mining Concept and Techniques" by Jiawei Han and Micheline Kamber,Publication-Elsevier,Second Edition 2005.*

[15] *"Data Mining Practical Machine Learning Tools and Techniques" by Ian H. Witten & Eibe Frank, Publication-Elsevier, Second Edition 2005.*