DATA MINING APPROACHES FOR NETWORK INTRUSION DETECTION SYSTEM

[1]T.S.Meenatchi, [2]K. Mythili, [3]M. Gayathri

Research Scholar, Assistant Professor, Assistant Professor
[1,2]CSA Department, [3]CSE Department
Sri Chandrasekharendra Saraswathi Viswa Mahavidhyalaya University, Kancheepuram, Tamil Nadu, India

***ABSTRACT*—With the rapidly development of internet, Network Security has become the key issue of network based services and information sharing on networks. Intruders are monitoring computer network continuously for attacks. Intrusion happens simply when the security and privacy of a system is compromised. To protect this, a sophisticated firewall with efficient intrusion detection system (IDS) is required to prevent computer network from attacks. Intrusion Detection System (IDS) plays very important role in network security as it detects various types of attacks in network. An effective Intrusion detection system requires high accuracy rate (True positive) and low false alarm (False Positive and False Negative) rate. In this paper data mining techniques SMO (Support Vector Machine), IBk (k-Nearest Neighbour), Attribute Selected classifier(Meta), J48 (Tree) and KDD99 data set is used to evaluate these Data mining algorithms which is most efficient and high accuracy rated.**

I. INTRODUCTION

As the cost of the information technology and Internet usage falls, societies are becoming wide variety of cyber threats. According to a recent survey, the rate of cyber attacks has been more than doubling every year in recent times. It has become increasingly important to make our information systems, especially in the defense, banking, commercial, public sectors.

Intrusion detection includes identifying a set of malicious actions that compromise the integrity, confidentiality, and availability of information resources. Data mining based intrusion detection techniques generally fall into one of two categories; misuse detection and anomaly detection. In misuse detection, each instance in a data set is labeled as 'normal' or 'intrusion' and a learning algorithm is trained over the labeled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labeled appropriately. A key advantage of misuse detection techniques is their high degree of accuracy in detecting known attacks and their variations. Anomaly detection techniques thus identify new types of intrusions as deviations from normal usage. While an extremely powerful and novel tool, a potential drawback of these techniques is the rate of false alarms. This can happen primarily because previously unseen system behaviors may also be recognized as anomalies, and hence flagged as potential intrusions.

Intrusion detection system collects online information from the network after that monitors and analyzes these information and partitions it into normal & malicious activities, provide the result to system administrator .

Intrusion Detection Using Data Mining Techniques

Intrusion detection is one of the major information security problems. IDS assist the system in resisting external attacks. Data mining techniques are applied on IDS because it can extract the hidden information and deals with large dataset. Presently Data mining techniques plays a vital role in IDS. By using Data mining techniques, IDS helps to detect abnormal and normal patterns. The following data mining techniques can be used in intrusion detection, each with its own specific advantage. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for Network intrusion detection.

Aim of Study

The aim of study deals with to determine the accuracy detection rate of Intrusion detection using data mining classification techniques .

Statement Of The Problem

The present study deals with the detection performance rate of various network intrusions using data mining approaches like Classification Algorithms, and functions using weka. Any of the existing commercial tool or technique is not fully detect the vast collection of intrusion methods.

Objectives Of Study

The objectives for the present study are as follows.

[1] To know the various Intrusion detection methods
[2] To know the data mining approaches for network intrusion detection.
[3] To evaluate the various intrusion detection tools using data mining.

[4] To normalize the KDDCup'99 dataset by using the rough set theory.
[5] To analyse the experimental results on the KDDCup'99 dataset have demonstrated that our intrusion detection models are much more efficient in the detection of intrusive behavior in standard classification techniques.

Scope of Study

Data mining is the analysis of dataset to find unsuspected relationship and to summarize large amounts of data in novel ways that are both understandable and useful to data owner in proactive decision making. It delivers new algorithms that can automatically sift deep into your data at the individual record level to discover patterns, relationships, factors, clusters, associations, profiles, and predictions that were previously "hidden" by using the techniques of data mining. To apply the techniques to information security we used a KDD 99 Cup datasets. The KDD 99 Cup consists of 42 attributes and 500,000 rows. To reduce the 500,000 dataset we use the Database Normalization technique. Dataset can be reduced by using the KDD 99 Cup Dataset Normalization Based on Rough Set Theory. In the present study deals with the accuracy level of various network intrusions based on the data mining classification techniques and the functions using weka. We analyse the KDD cup-99 Dataset for test and trained data of 50,000 instances to apply the weka tool to get the result of various classification algorithms and compare the results to obtain the accuracy level of best one.

Hypothesis of the Study

[1] There is Association Between the Intrusion Types and the Protocol type.
[2] J48 Classification tree is more efficient in case of known and unknown attacks.
[3] There is no performance difference between SMO (Sequential Minimal Optimization) and IBk- kNN Algorithms while detecting unknown intrusions.
[4] There is no performance difference between IBk – kNN and J48 Tree Classification algorithms in the domain of intrusion detection in high network traffic.

Method of Study

In the present study, the researcher employed Various Classification Algorithms including Function, Lazy, Meta, Decision Table and J48 Analysis to predict the Intrusion types in High Accuracy Level. For that purpose we use Weka 3.8.0 Machine Learning Software. We analysis the training and test datasets with 50000 instances using standard KDD CUP 99 Dataset.

II. LITERATURE REVIEW

The review of literature promotes a greater understanding of the problem and its crucial aspects and ensures the avoidance of unnecessary duplication. It also provides comprehensive data on the basis of which the researcher can evaluate and interpret the significance of his findings.

Sumaiya Thaseen et.al.(2013), analyzed different tree based classification techniques for IDS. Experimental results show that Random tree model reduces false alarm rate and has highest degree of accuracy.

Krishna Kant Tiwari et.al.(2010) observed that, most researches focus on anomaly detection, and use the tuples of DARPA1998 and KDDCup1999 mostly. In addition, most researches in intrusion detection use Artificial Neural Network(ANN). Because ANN is much more stable and reliable than other models and algorithms. Besides, the second most used model is Support vector machines (SVM).

Paul Dokas et.al., observed that, several intrusion detection schemes for detecting net-work intrusions are proposed in this paper. When applied to KDDCup'99 data set, developed algorithms for learn-ing from rare class were more successful in detecting network attacks than standard data mining techniques. Experimental results performed on DARPA 98 and real network data indicate that the LOF approach was the most promising technique for detecting novel intrusions. When performing experiments on DARPA'98 data, the unsupervised SVMs were very promising in detecting new intrusions but they had very high false alarm rate. Therefore, future work is needed in order to keep high detection rate while lowering the false alarm rate.

M.Govindarajan et.al.(2009), proposed new K-nearest neighbour classifier applied on Intrusion detection system and evaluate performance in term of Run time and Error rate on normal and malicious dataset. This new classifier is more accurate than existing K-nearest neighbour classifier.

R. China Appala Naidu et.al.(2012), used three Data mining techniques SVM, Ripper rule and C5.0 tree for Intrusion detection and also compared the efficiency. By experimental result, C5.0 decision tree is efficient than other. All the three Data mining technique gives higher than 96% detection rate.

Mohammadreza Ektela et.al.(2010), used Support Vector Machine and classification tree Data mining technique for intrusion detection in network. They compared C4.5 and Support Vector Machine by experimental result and found that C4.5 algorithm has better performance in term of detection rate and false alarm rate than SVM, but for U2R attack SVM performs better.

P. Amudha et.al.(2011), observed that Random forest gives better detection rate, accuracy and false alarm rate for Probe and DOS attack & Naive Bayes Tree gives better performance in case of U2R and R2L attack. Also the execution time of Naive Bayes Tree is more as compared to other classifier.

III. METHODOLOGY

Method Of Data Collection

The researcher collected the needed dataset from the KDD archive popularly referred to as the KDD 99 Cup dataset. The KDD 99 Cup consists of 42 attributes and 500,000 rows. We receive the dataset from the website of University of California, Irvine(UCI) Knowledge Discovery Databases(http://kdd.ics.uci.edu).

KDD CUP 99 Data Set

The KDD training dataset consist of 10% of original dataset that is approximately 494,020 single connection vectors each of which contains 42 features and is labeled with exact one specific attack type i.e., either normal or an attack. Each vector is labeled as either normal or an attack, with exactly one specific attack type. Deviations from 'normal behavior', everything that is not 'normal', are considered attacks. Attacks labeled as normal are records with normal behavior.

Normalization of Database

Database Normalization is a technique of organizing the data in the database. Normalization is a systematic approach of decomposing tables to eliminate data redundancy and undesirable characteristics like Insertion, Update and Deletion Anomalies. It is a multi-step process that puts data into tabular form by removing duplicated data from the relation tables. Normalization is used for mainly two purpose,

- Eliminating redundant (useless) data.
- Ensuring data dependencies make sense i.e data is logically stored.

Normalization rule are divided into following normal form.

1. First Normal Form
2. Second Normal Form
3. Third Normal Form
4. BCNF (Boyce and Codd Normal Form (BCNF)

KDD Cup99 Dataset Normalization Based on Rough Set Theory

Rough Sets

Rough sets have been introduced as a tool to deal with inexact, uncertain or vague knowledge in artificial intelligence applications. Here we use some basic notions related to information systems and rough sets.

An *information system* is a pair $\boldsymbol{A} = (U, A)$, where U is a non-empty, finite set called the *universe* and A - a non-empty, finite set of attributes, i.e. a: $U \rightarrow V_a$ for $a \in A$, where V_a is called the *value set* of a. Elements of U are called objects and interpreted as, for example, cases, states, processes, patients, observations. Attributes are interpreted as features, variables, characteristic conditions, etc.

Every information system $\boldsymbol{A} = (U, A)$ and non-empty set $B \subseteq A$ determine a B-information function

$$\boldsymbol{Inf_B: U \rightarrow P(B \times \bigcup_{a \in B} V_a)} \textbf{ defined by } \boldsymbol{Inf_B(x) = \{(a,a(x)): a \in B\}.}$$

We define B - indiscernibility relation as follows:

$$\boldsymbol{xIND(B)y \textbf{ iff } Inf_B(x)=Inf_B(y).}$$

For every subset $X \subseteq U$ we define the lower approximation **IND(B) (X)** and the upper approximation **IND (B) (X)** as follows:

$$\boldsymbol{IND\ (B)\ (X) = \{x \in U: [x]_B \subseteq X\},}$$
$$\boldsymbol{IND\ (B)\ (X) = \{x \in U: [x]_B \cap X \neq \theta}$$

After reducing the dataset we got a smaller version 10% training dataset for more memory required machine learning algorithms. The training dataset has 50,000 instances with 42 attributes. KDD CUP 99 has been most widely used in attacks on network.

Attacks Fall Into Four Categories

Denial Of Service Attack (Dos)

In this category the attacker makes some computing or memory resources too busy or too full to handle legitimate request, or deny legitimate users access to machine.

DOS contains the attacks: 'neptune', 'back', 'smurf', 'pod', 'land', and 'teardrop'.

Users to Root Attack (U2R)

In this category the attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to obtain root access to the system.

U2R contains the attacks: 'buffer_overflow', 'loadmodule', 'rootkit' and 'perl'

Remote to Local Attack (R2L)

In this category the attacker sends packets to machine over a network but who does not have an account on that machine and exploits some vulnerability to gain local access as a user of that machine.

R2L contain the attacks: 'warezclient', ' multihop', 'ftp_write', 'imap', 'guess_passwd', 'warezmaster', 'spy' and 'phf'.

Probing Attack(PROBE)

In this category the attacker attempt to gather information about network of computers for the apparent purpose of circumventing its security.

PROBE contains the attacks: 'portsweep', 'satan', 'nmap', and 'ipsweep'

The major objectives performed by detecting network intrusion are stated as recognizing rare attack types such as U2R and R2L, increasing the accuracy detection rate for suspicious activity, and improving the efficiency of real-time intrusion detection models. This detects that the training dataset consisted of 50,000 records, among which **97,277 (19.69%) were 'normal',**

391,458(79.24%) DOS, 4,107 (0.83%) Probe, 1,126 (0.23%) R2L and 52 (0.01%) U2R attacks. Each record has 42 attributes describing different features and a label assigned to each either as an 'attack' type or as 'normal'.

Protocols used in KDD :

TCP Protocol:

TCP stands for "Transmission Control Protocol". TCP is an important protocol of the Internet Protocol Suite at the Transport Layer which is the fourth layer of the OSI model . It is a reliable connection-oriented protocol which implies that data sent from one side is sure to reach the destination in the same order. TCP splits the data into labeled packets and sends them across the network. TCP is used for many protocols such as HTTP and Email Transfer.

UDP Protocol :

UDP stands for "User Datagram Protocol". It is similar in behavior to TCP except that it is unreliable and connection-less protocol. As the data travels over unreliable media, the data may not reach in the same order, packets may be missing and duplication of packets is possible. This protocol is a transaction-oriented protocol which is useful in situations where delivery of data in certain time is more important than loosing few packets over the network. It is useful in situations where error checking and correction is Possible in application level.

ICMP Protocol

Stands for "Internet Control Message Protocol". ICMP is basically used for communication between two connected computers. The main purpose of ICMP is to send messages over networked computers. The ICMP redirect the messages and it is used by routers to provide the up-to-date routing information to hosts, which initially have minimal routing information. When a host receives an ICMP redirect message, it will modify its routing table according to the message.

IV DATA ANALYSIS AND INTERPRETATION

Tools And Algorithms Used

The software tool and the effective algorithms are used to analyzing the KDD CUP99 dataset for high accuracy detection rate. For this purpose we use the standard Weka3.8 tool developed by University of Waikato, New Zealand. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Weka contains lot of classifier algorithms that are preloaded in weak explorer or an option to install more algorithms from package manager. The popular Classification algorithms are SMO, IBK, Attribute Selected Classifier , J48, .

SMO (Sequential Minimal Optimization Algorithm in Support Vector Machine)

Sequential minimal optimization (SMO) is an algorithm for solving the quadratic programming (QP) problem that arises during the training of support vector machines. It was invented by John Platt in 1998 at Microsoft Research. SMO is widely used for training support vector machines and is implemented by the popular LIBSVM tool. The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex and required expensive third-party QP solvers.

IBk (Instance Based K Nearest Neighbours Classifier)

K-nearest neighbour (*k*-NN) method assumes all instances correspond to points in the *n*-dimensional space. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance. *K-NN* algorithm is easily adopted to perform a real-valued prediction that is to approximate the continuous-valued target function. Instead of calculating the most common value of the nearest training instances, it has to calculate their mean value.

Attribute Selected Classifier (Meta)

Weka also offers a meta-classifier that takes a search algorithm and evaluator next to the base classifier. This makes the attribute selection process completely transparent and the base classifier receives only the reduced dataset.

J48 Tree (C4.5 Decision Tree)

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

V RESULT AND DISCUSSIONS

Table 1 Association Between the Intrusion types and the protocol type

S.No	Protocol Type	Attack Type	No of Attacks
1	TCP,UDP	Normal	34655
2	UDP	SNMPGetAttack	702
3	ICMP	Smurf	4102
4	TCP	Neptune	4738
5	TCP	Satan	1583
6	TCP	Mailbomb	111

7	ICMP	Saint	419
8	TCP	Guess_Passwd	447
9	ICMP	IPSweep	1152
10	TCP	PortSweep	1112
11	UDP	TearDrop	979
Total no of Attacks (including normal)			**50,000**

From the above table the Maximum number of Attacks are found in TCP protocol that mean 52% of attacks (7991/15345) are in TCP protocol. So the Hypothesis is Retained.

Table 2 PERFORMANCE METRICS

Classifier	Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Time Taken to Test Model	Accuracy Level
Function	SMO	49194	806	6.63 Seconds	98.388%
Lazy	IBk – kNN	49788	212	1392.77 seconds	99.576%
Meta	AttributeSelected Classifier	49346	654	3.1 Seconds	98.692%
Tree	J48	49712	288	1.61 Seconds	99.424%

The above table clearly concluded that, IBk –kNN and J48 Tree Classifiers are gives high prediction level with more accuracy.

ANALYSIS RESULT :

1. Maximum number of attacks like SNMPGetAttack, Neptune, Satan, Mailbomb, Guess_Password and Portsweep are found only in TCP protocol type.
2. Neptune and Smurf attacks are mostly hit the TCP and ICMP protocols.
3. All algorithms like SMO, IBk-kNN, AttributeSelectedClassifier, Decision Table and J48 Tree are gives fairly accurate results.
4. IBk-kNN gives the high level of accuracy (99.576%) and Low level of incorrectly classified instances (only 212) to compare with other Classifier algorithms. But the processing speed is very low that means it takes more time (1392.77 seconds) to test the model.
5. J48 gives fairly accurate level(99.424%) compared with IBk-kNN(99.576%). But the Processing speed is very high that means it takes very less seconds (1.61 seconds) to predict the model. It is most important feature, because the Network Intrusion Detection systems are functioning only in Real Time Applications like Servers etc..

VI CONCLUSION

J48 is the most powerful tool with high detection rate and 99.424% accuracy level. The Second most significant model is IBk-kNN which gives more accuracy(99.576%) than J48 Tree. But the drawback of kNN is its processing time. kNN take more time to execute the predictions. The SVM and ANN algorithms are also more stable and reliable like other models. The execution times of SVM and ANN are gives less processing time and high accuracy only in smaller datasets. But in large datasets like Real-time application, Network Tracking, Cloud Servers that is not suitable because of its extra processing time.The Decision tree algorithms (J48) has high detection rate with high accuracy in case of large dataset. Any single classifier alone is not sufficient to achieve high accuracy (100%) and low false positive or negative. Future work in this observation involves discovering new approach to improve the performance of network intrusion detection system using data mining.

REFERENCES

[1] W. Li, "Using Genetic Algorithm for Network Intrusion Detection". "A Genetic Algorithm Approach to Network intrusion Detection". SANS Institute, USA, 2004.
[2] W. Lu, I. Traore, "Detecting New Forms of Network Intrusion Using Genetic programming". Computational Intelligence, vol. 20, pp. 3, Blackwell Publishing, Malden, pp. 475-494, 2004.
[3] Bridges and Vaughn, Wang, Wendong, and Susan M. Bridges, "Genetic Algorithm Optimization of Membership Functions for Mining Fuzzy Association Rules," Proceedings of the 7th International Conference on Fuzzy Theory & Technology, Atlantic City, NJ, February 27 – March 3, 2000, pp.131-134.
[4] Bridges, Susan M., and Rayford M. Vaughn, "Fuzzy Data Mining and Genetic Algorithms Applied to Intrusion Detection," Proceedings of the Twenty-third National Information Systems Security Conference, Baltimore, MD, October 2000.
[5] Xia et al, Tao Xia, Guangzhi Qu, Salim Hariri and Mazin Yousif, "An efficient Network Intrusion Detection Method Based on Information Theory andGenetic Algorithm", IEEE, 2005.
[6] Li, w. Using genetic algorithm for network intrusion detection. Proceedings of the united states department of energy cyber security group 2004 training conference, may 24 27, kansas city, ks
[7] Crosbie and spafford Crosbie, Mark, and Gene Spafford. 1995. "Applying Genetic Programming to Intrusion Detection." In Proceedings of 1995 AAAI Fall Symposium on Genetic Programming , pp. 1-8. Cambridge, Massachusetts.

[8] Xiang m.y. chang et.al,Xiang, M.Y. Chong and H. L. Zhu, "Design of Multiple - level Tree classifiers for intrusion detection system ", IEEE conference on Cybernetics and Intelligent system, 2004
[9] Song naiping et.al,Song Naiping and Zhou Genyuan, "A study on Intrusion Detection Based on Data Mining", International Conference of Information Science and Management Engineering , Pp 135-138, IEEE,2010
[10] Deepthy k denatious et.al Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", International Conference on Computer Communication and informatics, 2012.