A Comprehensive Analysis and study in Intrusion Detection System using Data Mining Techniques

G.V. Nadiammai
Ph.D (CS) Research Scholar
Karpagam University
Coimbatore - 641 021

S.Krishnaveni
Ph.D (CS) Research Scholar
Karpagam University
Coimbatore - 641 021

M. Hemalatha
Head, Dept. Software
Systems & Research,
Karpagam University
Coimbatore - 641 021

ABSTRACT

Data mining refers to extracting knowledge from large amounts of data. Most of the current systems are weak at detecting attacks without generating false alarms. Intrusion detection systems (IDSs) are increasingly a key part of system defense. An intrusion can be defined as any set of actions that compromise the integrity, confidentiality or availability of a network resource(such as user accounts, file system, kernels & so on).Data mining plays a prominent role in data analysis. In this paper, classification techniques are used to predict the severity of attacks over the network. I have compared zero R classifier, Decision table classifier & Random Forest classifier with KDDCUP 99 databases from MIT Lincoln Laboratory.

Keywords

Data Mining, Intrusion Detection, Machine Learning, Zero R, Decision Table & Random Forest classifier, KDDCup99 dataset

1. INTRODUCTION

1.1 Data mining

Data mining [4] is also known as knowledge discovery in databases has attained a great deal of attention in the information industry & in society. Within few years, the availability of large amount of data & its prominent need for extracting such data into useful information is increasing rapidly. Various machine learning algorithms, for instance Neural Network, Support Vector Machine, Genetic Algorithm, Fuzzy Logic, and Data Mining have been extensively used to detect intrusion activities both for known and unknown dynamic datasets.

Data mining tasks can be classified into 2 categories namely descriptive mining & predictive mining. The descriptive mining techniques such as clustering, Association, Sequential Pattern discovery, is used to find human interpretable patterns that describe the data. The predictive mining techniques like classification, Regression, and Deviation detection, etc., are used to predict unknown or future values of other variables.

1.2 IDS

Intrusion activities to computer systems are increasing due to the commercialization of the internet & local networks. An intrusion detection system watches networked devices & searches for malicious behaviors in kinds of pattern in the audit stream [13]. One main conflict in intrusion detection [15] is that we have to find out the hidden attacks from a large quantity of routine communication activities. The security of our computer systems & data is at continual risks due to the extensive growth of the internet & increasing availability of tools & tricks for intruding & attacking networks have made intrusion detection to become a critical component of network administration.

1.3 Types of IDS

a) Host-based IDS

Host based IDSs examine data held on individual computer that serve as hosts. The network structural design of host based is an agent-based, which means that software resides on each of the hosts that will be governing by the system

b) Network-based IDS

Network based IDSs analyses data exchanged between computers. Most efficient host-based intrusion detection systems [27] are capable of monitoring and gathering system audit in real time as well as on a scheduled basis, thus by utilizing both CPU utilization and network. It also provides a flexible means of security administration. Each technique has a unique approach for monitoring and securing data and each group has its own advantages and disadvantages. During IDS implementation, it is better to incorporate the network intrusion detection system to filter alerts and notifications in the host based system, controlled from the same central location. This provides a convenient means of managing and acting against misuse by using both types of intrusion detection.

1.4 Detection Approaches

a) Misuse Detection

It searches for patterns or user behavior that matches known intrusion or scenarios, which are stored as signatures. These hand coded signatures are laboriously provided by human experts based on their knowledge. If a pattern match is found, it signals an event then an alarm is raised. But it is unable to detect new or previously unknown intrusion.

b) Anomaly Detection

Anomaly detection includes profiles (normal network behavior) which can be used to detect new patterns that deviate from the profiles. The main advantage of anomaly detection is that it may detect new intrusion that have not yet observed. A limiting factor of anomaly detection is the high percentage of false positives.

2. RELATED WORK

Intrusion detection concept was introduced by **James Anderson** in 1980[4] defined an intrusion attempt or threat to be potential possibility of a deliberate unauthorized attempt to access information, manipulate or render a system unreliable or unusable. Sights moved for using data mining in content of NIDS in the late of 1990's. Researchers suddenly recognized the need for existence of standardized dataset to train IDS tool. Minnesota Intrusion Detection System (MINDS) combines signature based tool with data mining techniques. Signature based tool (Snort) are used for misuse detection & data mining for anomaly detection.

In [12] **Jake Ryan et al** applied neural networks to detect intrusions. Neural network can be used to learn a print (user behavior) & identify each user. If it does not match then the system administrator can be alerted. A back propagation neural network called NNID was trained for this process.

Denning D.E et al [6] has developed a model for monitoring audit record for abnormal activities in the system. Sequential rules are used to capture a user's behavior [26] over time. A rule base is used to store patterns of user's activities deviates significantly from those specified in the rules. High quality sequential patterns are automatically generated using inductive generalization & lower quality patterns are eliminated. An automated strategy for generation of fuzzy rules obtained from definite rules using frequent items. The developed system [21] achieved higher precision in identifying whether the records are normal or attack one.

Dewan M et al [7] presents an alert classification to reduce false positives in IDS using improved self adaptive Bayesian algorithm (ISABA). It is applied to the security domain of anomaly based network intrusion detection.

S.Sathyabama et al [20] used clustering techniques to group user's behavior together depending on their similarity & to detect different behaviors and specified as outliers.

Amir Azimi Alasti et al [3] formalized SOM to classify IDS alerts to reduce false positive alerts. Alert filtering & cluster merging algorithms are used to improve the accuracy of the system.SOM is used to find correlations between alerts.

Alan Bivens et al [1] has developed NIDS using classifying self organizing maps for data clustering. MLP neural network is an efficient way of creating uniform, grouped input for detection when a dynamic number of inputs are present.

An ensemble approach [24] helps to indirectly combine the synergistic & complementary features of the different learning paradigms without any complex hybridization. The ensemble approach outperforms both SVMs MARs & ANNs. SVMs outperform MARs & ANN in respect of Scalability, training time, running time & prediction accuracy. This paper [23] focuses on the dimensionality reduction using feature selection. The Rough set support vector machine (RSSVM) approach deploy Johnson's & genetic algorithm of rough set theory to find the reduct sets & sent to SVM to identify any type of new behavior either normal or attack one.

Aly Ei-Senary et al [2] has used data miner to integrate Apriori & Kuok's algorithms to produce fuzzy logic rules that captures features of interest in network traffic.

Taeshik Shon et al [25] proposed an enhanced SVM approach framework for detecting & classifying the novel attacks in network traffic. The overall framework consist of an enhanced SVM- based anomaly detection engine & its supplement components such as packet profiling using SOFM, packet filtering using PTF, field selection using Genetic Algorithm & packet flow-based data preprocessing. SOFM clustering was used for normal profiling. The SVM approach provides false positive rate similar to that of real NIDSs. In this paper [19] genetic algorithm can be effectively used for formulation of decision rules in intrusion detection through the attacks which are more common can be detected more accurately.

Oswais.S et al [18] proposed genetic algorithm to tune the membership function which has been used by IDS. A survey was performed using approaches based on IDS, and on implementing of Gas on IDS.

Norouzian M.R et al [17] **defined** Multi- Layer Perceptron (MLP) for implementing & designing the system to detect the attacks & classifying them in six groups with two hidden layers of neurons in the neural networks. Host based intrusion detection is used to trace system calls. This system [21] does not exactly need to know the program codes of each process. Normal & intrusive behavior are collected through system call & analysis is done through data mining & fuzzy technique. The clustering and genetic optimizing steps [14] were used to detect the intrude action with high detection rate & low false alarm rate.

3. CLASSIFICATION MODEL DESCRIPTION

• Zero R Classifier

Zero R is the simplest classification method which is place on the target and ignores all predictors. Zero R classifier simply predicts the majority class. There is no predictability control in Zero R and it is useful for making a baseline performance as a standard for other classification methods.

• Decision Table Classifier

Decision table uses simple Boolean values to represent the alternatives to a condition, other tables use numbers, and some tables also use fuzzy logic or probabilistic representations for condition alternatives. In same way an action entry can simply represent whether an action is to be performed, or in more advanced decision tables, the sequencing of actions to perform.

• Random Forest Classifier

Random Forest was formulated by Tin Kam Ho of Bell Labs in 1995. This method combines bagging and the random selection of features to construct a group of decision trees with controlled variation. The selection of a random subset of features is a method of random subspace method, which is a way to implement stochastic bias proposed by Eugene Kleinberg.

4. KDD CUP 99 Dataset

The 1998 DARPA Intrusion Detection Evaluation Program was prepared and managed by MIT Lincoln Labs. The objective was to survey and evaluate research in intrusion detection. A standard set of data includes a wide variety of intrusions simulated in a military network environment [11]. The DARPA 1998 dataset includes training data with seven weeks of network traffic and two weeks of testing data providing two million connection records. A connection is a sequence of TCP packets starting and ending at some well defined times, between source IP address to a target IP address with some well defined protocol. Each connection is categorized as normal, or as an attack, with one specific attack type.

The training dataset is classified into five subsets namely **Denial of service attack, Remote to Local attack, User to Root attack, Probe attacks and normal data**. Each record is categorized as normal or attack, with exactly one particular attack type.

Table 1: Various types of attacks described in four major categories

Denial of Service Attacks	Back, land, neptune, pod, smurf, teardrop
Probes	Satan, ipsweep, nmap, portsweep
Remote to Local Attacks	Ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
User to Root Attacks	Buffer_overflow, load module, Perl, root kit

• Denial of Service Attacks:

In denial of service the attacker develops some computing or memory resource available or unavailable to manage valid requirements, or reject valid user's rights to use a machine.

• User to Root Attacks:

In User to Root [16] attack, the attacker initiate by using a normal user account on the system and take advantage of some vulnerability to achieve root access to the system.

• Remote to User Attacks:

In Remote to User attack takes place when an attacker has the ability to send packets to a machine over a network but does not have an account on that machine, performing some vulnerability to access as a user of that machine.

• Probes:

Probing is a kind of attacks takes place when an attacker checks a network to collect information or find out well-known threats. This information is helpful for an attacker who is plans to make an attack in future. There are different types of probes such as abusing the system's legitimate features, using social engineering methods. However this type of attack requires few technical expertises.

5. RESULTS & DISCUSSION

This work deals with the performance of three classification algorithms namely Zero R, Decision Table & Random Forest classifiers.Kddcup99 dataset produced by Lincoln Laboratory at MIT where each record has been specified as normal or attacked one with specific type of attacks. The dataset is divided into four different scenarios:

A. **Based on Dos**
B. **Based on Probe**
C. **Based on R2L**
D. **Based on U2R**

Table 2: Correctly (Cc) and Incorrectly (Icc) Classified sample of Kddcup99 dataset

Classifiers		**Dos**		**Probe**	
		#Record	**Accuracy%**	**#Record**	**Accuracy%**
Zero R	Cc	3349	38.78	1589	38.78
	Icc	5294	61.22	2509	61.22
Decision Table	Cc	7325	94.50	2959	72.21
	Icc	1318	15.25	1139	27.79
Random Forest	Cc	7220	90.59	2858	69.74
	Icc	1423	16.25	1240	30.26

Classifiers		**R2L**		**U2R**	
		#Record	**Accuracy%**	**#Record**	**Accuracy%**
Zero R	Cc	1020	88.59	20	57.70
	Icc	106	11.41	32	42.30
Decision Table	Cc	1064	84.75	50	85.12
	Icc	62	5.50	2	14.88
Random Forest	Cc	1020	83.75	45	81.45
	Icc	106	9.41	7	18.55

Table 3: Attack Dataset Classification

Classifier	Dos	Probe	R2L	U2R
Zero R	38.78	38.78	88.59	57.70
Decision Table	94.50	72.21	84.75	85.12
Random Forest	90.59	69.74	83.75	81.45

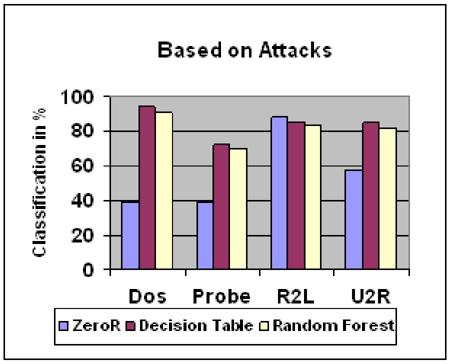

Figure 1: Comparison of Zero R, Decision Table and Random Forest based on attack dataset

Above figure 1 shows the attack dataset graph. In this Dos attribute, the accuracy of Zero R is 38.78%, for decision table the accuracy is 94.50% & for Random Forest the accuracy is 90.59%. For Probe, the accuracy of Zero R is 38.78%, the accuracy of decision Table is 72.21% & Random Forest the accuracy is 69.74%. In R2L attribute, the correctly classified Percentage is 88.59, for Decision table the accuracy is 84.75% & for Random Forest the accuracy is 83.75%. In U2R attribute, the accuracy of Zero R is 57.70%, the accuracy of Decision Table is 85.12% & in Random Forest the accuracy is 81.45%. Among these, the Random classification algorithm took highest percentage when compared with other classification algorithms.

A. Based on Dos Attacks

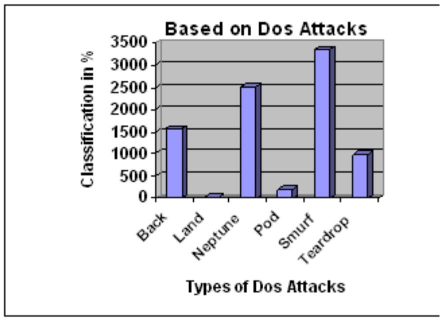

Figure 2 Dos attack type

Above figure shows the Dos attack graph. According to kddcup99 dataset Dos includes six types of attacks namely Back, Land, Neptune, Pod, and Smurf & Teardrop. Out of 8643 records in dos, back type includes 1570, land includes 21 records, Neptune includes 2512, and pod 210 & smurf includes 3348 records. Finally Smurf attack is found to be more when compared to other attacks totally.

B. Based on Probe Attacks

According to the figure 3 Probe attack includes four types of attacks based on kddcup99 databases. They are Nmap, Port sweep, Ipsweep & Satan. Out of 4098 records in probe, ipsweep includes 1248, Nmap includes 221, port sweep includes 1039 & Satan includes 1588 totally. Finally Satan attack is found to be more when compared to other attacks totally.

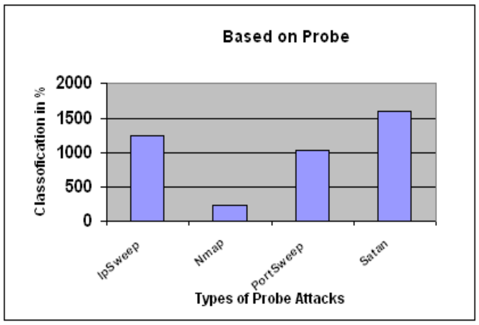

Figure 3 Probe attack type

C. Based on U2R Attacks

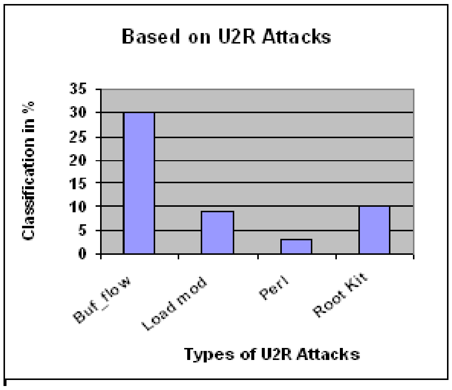

Figure 4 U2R attack type

The Figure 4 shows the U2R attack graph. The U2R attacks include five types of attacks namely Buffer Overflow, Load Module, Perl and Root kit respectively. Out of 52 records in U2R, buffer overflow includes 30, load module includes 9, Perl includes 3 & root kit includes 10 records. Finally buffer overflow attack is found to be more when compared to other attacks totally.

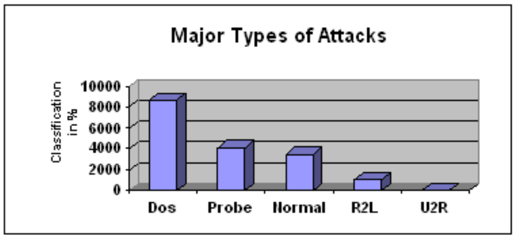

Figure 5 the sample dataset

In the above figure, major types of attacks are listed according to the dataset which I have taken for my work based on kddcup99 databases.

6. CONCLUSION

The aim of this paper is to detect the severity of attacks in the dataset based on kddcup99 dataset produced by MIT Lincoln Laboratory. With the help of this, the performance of Zero R, Decision Table & Random Forest classifiers are used to predict the classification accuracy. Based on this, Random Forest outperforms than other classification algorithms.

While comparing traditional intrusion detection systems, intrusion detection systems based on data mining are generally more precise & require less manual processing & input from human experts.

7. ACKNOWLEDGEMENT

We thank the Karpagam University and our guide Dr.M.Hemalatha for the Motivation and Encouragement to make this research work as successful one.

8. REFERENCES

[1] Alan Bivens, Chandrika Palagiri, Rasheda Smith, Boleslaw Szymanski, ”Network-Based Intrusion Detection Using Neural Networks”, in Proceedings of the Intelligent Engineering Systems Through Artificial Neural Networks, St.Louis, ANNIE-2002, and Vol: 12, pp- 579-584, ASME Press, New York.

[2] Aly Ei-Semary, Janica Edmonds, Jesus Gonzalez-Pino, Mauricio Papa, “Applying Data Mining of Fuzzy Association Rules to Network Intrusion Detection”, in the Proceedings of Workshop on Information Assurance United States Military Academy 2006, IEEE Communication Magazine, West Point, NY,DOI:10.1109/IAW.2006/652083.

[3] Amir Azimi, Alasti, Ahrabi, Ahmad Habibizad Navin, Hadi Bahrbegi, “A New System for Clustering & Classification of Intrusion Detection System Alerts Using SOM”, International Journal of Computer Science & Security, Vol: 4, Issue: 6, pp-589-597, 2011.

[4] Anderson.J.P, “Computer Security Threat Monitoring & Surveilance”, Technical Report, James P Anderson co., Fort Washington, Pennsylvania, 1980.

[5] Data Mining:Concepts and Techniques, 2nd Edition , Jiawei Han and Kamber,Morgan kaufman Publishers, Elsevier Inc,2006.

[6] Denning .D.E, ”An Intrusion Detection Model”, Transactions on Software Engineering, IEEE Communication Magazine, 1987,SE-13, PP-222-232,DOI:10.1109/TSE.1987.232894.

[7] Dewan Md, Farid, Mohammed Zahidur Rahman, “Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm”, Journal of Computers, Vol 5, pp-23-31, Jan 2010, DOI:10.4.304/jcp 5.1.

[8] ZeroR, available at http://en.Wikipedia.org/wiki/ZeroR

[9] Decision tree, available at http://en.Wikipedia.org/wiki/Decision_tree

[10] Random Forest, available at http://en.Wikipedia.org/wiki/Random_Forest

[11] KDD Cup 1999 Data, available at http://kdd.ics.uci.edu/databases/kddcup99/kdd cup99.html.

[12] Jake Ryan, Meng - Jang Lin, Risto Miikkulainen, ”Intrusion Detection With Neural Networks”, Advances in Neural Information Processing System 10, Cambridge, MA:MIT Press,1998,DOI:10.1.1.31.3570.

[13] Jian Pei, Upadhayaya.S.J, Farooq.F, Govindaraju.V,”Data Mining for Intrusion Detection: Techniques, Applications & Systems, in the Proceedings of 20th International Conference on Data Engineering, pp-877-887, 2004.

[14] Jin-Ling Zhao, Jiu-fen Zhao ,Jian-Jun Li, “Intrusion Detection Based on Clustering Genetic Algorithm”, in Proceedings of International Conference on Machine Learning & Cybernetics (ICML),2005, IEEE Communication Magazine,ISBN:0-7803-9091-1,DOI: 10.1109/ICML.2005.1527621.

[15] Macros .M. Campos, Boriana L. Milenora, “ Creation & Deployment of Data Mining based Intrusion Detection Systems in Oracle Db 10g”, in the proceedings of 4th International Conference on Machine Learning & Applications, 2005.

[16] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", in Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, pp. 53-58, Ottawa, Ontario, Canada, 2009.

[17] Norouzian.M.R, Merati.S, “Classifying Attacks in a Network Intrusion Detection System Based on Artificial Neural Networks”, in the Proceedings of 13th International Conference on Advanced Communication Technology(ICACT), 2011,ISBN:978-1-4244-8830-8,pp-868-873.

[18] Oswais.S, Snasel.V, Kromer.P, Abraham. A, “Survey: Using Genetic Algorithm Approach in Intrusion Detection Systems Techniques”, in the Proceedings of 7th International Conference on Computer Information & Industrial Management Applications (CISIM), 2008,

IEEE Communication Magazine,pp-300-307,ISBN:978-0-7695-318-7,DOI:10.1109/CISM.2008-49.

[19] Sadiq Ali Khan, "Rule-Based Network Intrusion Detection Using Genetic Algorithm", International Journal of Computer Applications, No: 8, Article: 6, 2011, DOI: 10.5120/2303-2914.

[20] Sathyabama.S, Irfan Ahmed.M.S, Saravanan.A,"Network Intrusion Detection Using Clustering: A Data Mining Approach", International Journal of Computer Application (0975-8887), Sep-2011, Vol: 30, No: 4, ISBN: 978-93-80864-87-5, DOI: 10.5120/3670-5071.

[21] Sekeh.M.A,Bin Maarof.M.A, "Fuzzy Intrusion Detection System Via Data Mining with Sequence of System Calls", in the Proceedings of International Conference on Information Assurance & security (IAS)2009,IEEE Communication Magazine, pp- 154-158,ISBN:978-0-7695-3744-3,DOI:10.1109/IAS.2009.32.

[22] Shanmugavadivu .R, "Network Intrusion Detection System Using Fuzzy Logic", Indian Journal of Computer Science & Engineering, and ISSN: 0976-5166, Vol: 2, No.1, pp- 101-110, 2011.

[23] Shilendra Kumar, Shrivastava ,Preeti Jain, "Effective Anomaly Based Intrusion Detection Using Rough Set Theory & Support Vector Machine(0975-8887), Vol:18,No:3, March 2011,DOI: 10.5120/2261-2906.

[24] Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham, "Intrusion Detection Using an Ensemble of Intelligent Paradigms",Journal of Network & Computer Applications ,pp-1-15, 2004.

[25] Taeshik Shon, Jong Sub Moon, "A Hybrid Machine Learning Approach to Network Anomaly Detection", Information Sciences 2007, Vol: 177, Issue: 18, Publisher: USENIX Association, pp- 3799-3821, ISSN:00200255,DOI:10.1016/j.ins-2007.03.025.

[26] Teng.H.S, Chen.K and Lu.S.C, "Adaptive Real-Time Anomaly Detection using Inductively Generated Sequential Patterns, in the Proceedings of Symposium on research in Computer Security & Privacy, IEEE Communication Magazine,1990, pp-278-284.

[27] Vera Marinova-Boncheva, "A Short Survey of Intrusion Detection Systems", Institute of Information Technologies, 1113 Sofia, pp-23-30, 2007.

8. AUTHORS PROFILE

Dr.M.Hemalatha completed MCA, M. Phil., Ph.D. in Computer Science and currently working as a Head, Dept. of software systems in Karpagam University. Ten years of experience in teaching and published sixty paper in International Journals and also presented seventy papers in various National and international conferences. Area of research is Data mining, Software Engineering, Bioinformatics and Neural Network also reviewer in several National and International journals.

G.V.Nadiammai completed MCA and currently pursuing Ph.D in computer science at Karpagam University under the guidance of Dr.M.Hemalatha, Head, Dept. of Software Systems & Research, Karpagam University, Coimbatore. Two papers published in International Journal. Area of Research is Data Mining.

S.Krishnaveni completed MCA, M. Phil., and currently pursuing Ph.D. in computer science at Karpagam University under the guidance of Dr.M.Hemalatha, Head, Dept. of Software Systems & Research, Karpagam University, Coimbatore. Three papers published in International Journal. Area of Research is Data Mining.