An Optimised Approach for Intrusion Detection in KDD CUP 99 Dataset Using KNN & GA

Megha Jain Gowadiya

Abstract— Security of the computer networks becomes tedious assignment due to the pervasive expansion in the utilization of it. Most of the person uses the network for sharing their private information such as messages, accounting information etc. and accessing the resources. But most of the happenstances possible when they are traveling over network from source to destination. To safeguard the network from such serious threats a system has been designed which is called intrusion detection system (IDS). These systems continuously monitor the actions performed on network and if they found any such malevolent activity or unauthorized venture it impedes them. Numerous data mining techniques have been used for designing efficient intrusion detection system such as support vector machine, Bayesian network, KNN classifier etc. In this work, we propose a modified data mining classification technique which helps to improve higher intrusion detection rate. The experimental analysis of the proposed system is perform using the function of MATLAB2012a toolbox and performance measurement is done using some metric such as accuracy. The simulation result of the proposed system for the accuracy parameter is improved than the existing system is about 5 %.

Index Terms — Accuracy, Data Mining, Intrusion detection, MATLAB, KDD CUP 99 Dataset, KNN, GA.

1 INTRODUCTION

THE The swift progress and widespread vogue nature of internet is resulted in to the next level security threats of networks. Safeguarding computer systems and networks worth huge significance and it has been a matter of great focus in current research scenario. The internet keeps growing with an exponential pace, so also is cyber attacks by crackers discovering loopholes in Internet protocols, operating system and application software. Several protective measures such as firewall have been put in place to check the activities of intruders which could not guarantee the full protection of the system. Hence, the needs for a more dynamic mechanism like intrusion detection system (IDS) as a second line of defense. Intrusion detection is the enterprise of monitoring events happening in a computer system or network and analyzing them for signs of intrusions. Intrusion detection systems can be classified into two categories based on the technique used to detect intrusions: anomaly detection and misuse detection [1], [2], [3]. Anomaly detection approach generates the profiles of normal activities of users, operating systems, system resources, network traffic and services using the audit trails generated by a host operating system or a network scanning program. This approach determines intrusions by identifying significant deviations from the normal behavioral patterns of these profiles. The strength of Anomaly detection approach is that prior knowledge of the security breaches of the target systems is not required. So, it is eligible to detect not only well known intrusions but also newbie unknown intrusions. In addition, this approach can detect the intrusions that are accomplished by the abuse of genuine users or masquerades without breaking security policy [4], [5]. The cons of this procedure were it had high rate of false positive detection error, the difficulty of handling gradual misbehavior, and expensive computation. In this work, we propose a data mining approach for the detection of network intrusions and the analysis of this approach is done using KDDCUP'99 dataset.

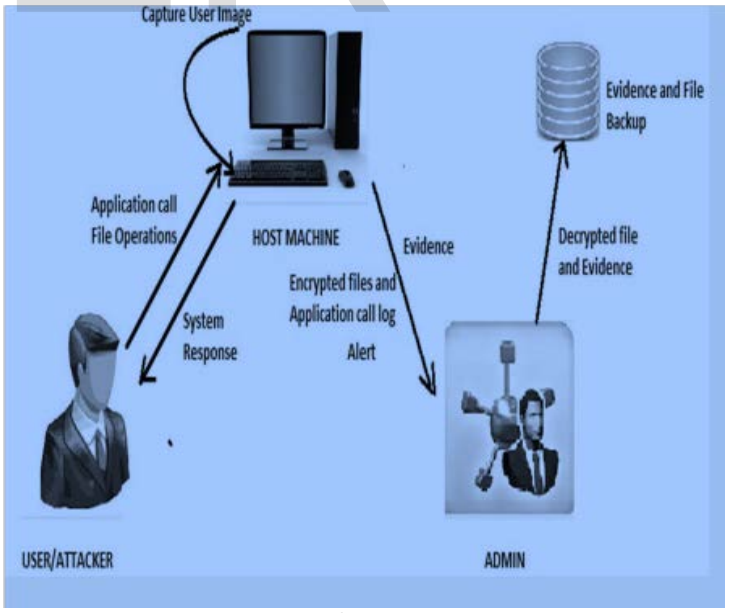

Figure 1.1 Architecture of Intrusion Detection System

2 SECTIONS

The organization of the remaining section of research paper is done as follows:

• *Megha Jain Gowadiya is currently pursuing masters degree program in computer science and engineering in Rajeev Gandhi Technical Univercity Bhopal, India*
E-mail: gmeghajain@gmail.com

Section III presents the former work done by various researchers for detection of intrusion.
Section IV describes the KDDCUP'99 test dataset. In section V describes our proposed approach to discover the novel threat which compromised from the networks. Section VI shows the experimental results and comparative analysis between propose and existing system. Last section presents the overall conclusion of our propose approach which is more efficient than existing.

3 RELATED WORK

Divyatmika et al. [6] presented a novel approach to build a network based intrusion detection system using machine learning approach. They proposed two-tier architecture to detect intrusions on network level. Network behavior can be classified as misuse detection and anomaly detection. If the analysis depends on the network behavior, they considered data packets of TCP/IP as input data. After, pre-processing the data by parameter filtering, they built an autonomous model on training set using hierarchical agglomerative clustering. Further, data gets filtered as regular traffic pattern or intrusions using KNN classification. This lowers cost-overheads. Misuse detection is conducted using MLP algorithm. Anomaly detection is conducted using Reinforcement algorithm where network agents learn from the environment and take decisions accordingly. The TP rate of our architecture is 0.99 and false positive rate is 0.01. Thus, our architecture provides a high level of security by providing high TP and low false positive rate. And, it also analyzes the usual network patterns and learns incrementally (to build autonomous system) to separate normal data and threats.

Laskov et al. [7] developed an experimental framework for comparative analysis of supervised (classification) and unsupervised learning (clustering) techniques for detecting ambivalent activities. They used two scenarios to evaluate the learning from both categories. They took training and test data from the same unknown distribution. The second scenario is based on the new or unseen data/patterns. This helps us understand how IDS can generalize its knowledge to new malicious patterns, which is often very essential for an IDS system.

Wankhede et al. [8] presented an association rule mining technique for IDS. Association mining was applied in order to generate the frequent patterns for various known attacks. These frequent patterns provide a baseline for intercepting the attacks from entering the system and also distinguish attacks from normal data, thus licensing normal data to enter the system.

Panda et.al, [9] proposed hybrid intelligent decision technologies using data filtering by adding guided learning methods along with a classifier to make more classified decisions in order to detect network attacks. It is noticed from the results found that the Naive Bayes model is quite appealing because of its integrity, elegance, robustness and effectiveness. However, decision trees have proven their efficiency in both generalization and detection of new attacks. The results show that there is no single best algorithm to outperform others in all situations. In certain cases there might be dependency on the characteristics of the data. To choose a worthy and suitable algorithm, a domain expert or expert system may employ the outcomes of the classification in order to fabricate better decisions.

Hemalatha et al. [10] data mining concept is integrated with IDS to identify the relevant, hidden data of interest for the user effectively and with less execution time. Four issues viz. Classification of Data, High Level of Human Interaction, Lack of Labeled Data, and Effectiveness of Distributed Denial of Service Attack are being solved using the proposed algorithms like EDADT algorithm, Hybrid IDS model, Semi-Supervised Approach and Varying HOPERAA Algorithm respectively. Our proposed algorithm has been tested using KDD Cup dataset. The entire proposed algorithm showed better accuracy and reduced false alarm rate when compared with existing algorithms.

Hassanat et al. [11] presented a new dataset is collected because there were no common data sets that contain modern DDoS attacks in different network layers, such as (SIDDoS, HTTP Flood). This work incorporates three well-known classification techniques: Multilayer Perceptron (MLP), Naïve Bayes and Random Forest. The experimental results show that MLP achieved the highest accuracy rate (98.63%).

Norouzian et al. [12] presented a most effective classification technique for detecting and classifying attacks into two groups normal or threat. They proposed a new approach to IDS based on a MultiLayer Perceptron Neural Network to detect and classify data into 6 groups. They implemented their MLP design with two hidden layers of neurons and achieved 90.78% accuracy rate.

Dong et al. [13] investigated the intrusion detection problem of the network defense, pointing at the problem of low fitting defects in the traditional detection algorithm of high precision and low forecasting accuracy under the situation of small sample training, and puts forward the algorithm of Support Vector Machine. Aimed at the important influence of SVM kernel function on classification performance, this paper adopts the improved Ant Colony Algorithm (ACA)as the method of selection SVM characteristics parameters. Experiments show this algorithm is significantly higher than the other algorithm in training and the detection speed, and have a high enhance of the detection rates of attacking sample.

Barakat et al. [14] new feature selection model is proposed; this model can effectively select the most relevant features for intrusion detection. Our goal is to build a lightweight intrusion detection system by utilising a reduced features set. Deleting irrelevant and redundant features helps to build a faster training and testing process, to have less resource consumption as well as to maintain high detection rates. The effectiveness and the feasibility of our feature selection model were

verified by several experiments on KDD intrusion detection dataset. The experimental results strongly showed that their model is not only able to yield high detection rates but also to rapidify the detection process

4 KDD CUP'99 DATASET

In the year 1998, the Defense Advanced Research Projects Agency (DARPA) intrusion detection estimation created the first standard compilation for evaluating intrusion detection systems. The offline intrusion detection 1998 evaluation was the first in a intended series of yearly assessment accomplished by the Massachusetts Institute of Technology (MIT) Lincoln Laboratories under DARPA protection. For calculating together counterfeit alarm rates and revealing rates of intrusion detection systems was deliberated by corpus using many types of in cooperation with known and novel attacks enclosed in

A large sum of normal surroundings traffic. More than 300 attacks were included in the 9 weeks of data collected for the evaluation. These 300 attacks were strained from 32 dissimilar attack types and 7 dissimilar attack scenarios as publicized in KDD dataset. Initial observations of the evaluation results for the 1998 competition concluded that most IDSs can simply recognize older, known attacks with a low false-alarm rate, even though do not execute as well when identifying novel or new attacks. Numerous extra intrusion detection challenges, such as DARPA 1999 and KDD Cup 1999, used related data sets to calculate results in intrusion detection investigation. The DARPA 1999 evaluation used a related structure for the competition, but incorporated Widows NT workstations in the simulation network. These evaluations of developing technologies are essential to focus endeavor manuscript existing potential and guide investigation. The DARPA evaluation focused on the development of evaluation corpora that could be used by many researchers for system designs and improvement. The estimation used the Receiver Operating Characteristic (ROC) method to calculate intrusion detection systems. The ROC approach analyzed the tradeoffs amongst false alarm rates and detection rates for detection method. ROC curves for intrusion detection designate how the detection rate changes as internal thresholds are varied to generate fewer more or fewer false alarms to tradeoffs detection accuracy against analyst workload. The training data was concerning four gigabytes of squashed binary transmission control protocol abandon data from seven weeks of network traffic which was processed into concerning five millions of group records. Also, the two weeks of test data consent around two million organized records. Accomplishing focused on the systems capability to detect novel attacks in the test data that was a variant of a known attack labeled in the training data. The KDD 99 training datasets enclosed overall of 24 training attack categories, with a supplementary 14 attack category in the test data merely [15]. The contributors were given a directory of high-level features that could be used to differentiate normal relations from attacks. An association is a progression of TCP packets starting and ending at an only some well defined times, surrounded by which data stream from a source IP address to a goal IP address under only some well defined protocols. Every association is labeled as either ordinary or as an attack with precisely one explicit attack type. All company manifestation incorporated of about 100 bytes. Three sets of feature were made obtainable for analysis. Initially, the same host characteristics scrutinize merely the links in the precedent two seconds that have the indistinguishable destination host as the current link, and conclude statistics related to protocol activities, provision, etc. The analogous identical service features inspect only the relations in the last two seconds which have the identical service as the present link. The identical host and identical service characteristics are jointly called time based traffic features of the link records. Various probing attacks scrutinize the hosts using a much more time period than two seconds, e.g. one time per minute. Consequently, supplementary features can be assembled using a window of 100 links to the same host as an alternative of a time window. This accepts a group of so called host based traffic features. Ultimately, domain knowledge can be used to engender features that appear for anxious behavior in the data segment such as the amount of failed login effort. These characteristics are called content characteristics. The networking attacks fall into four major categories [16]:

A. Denial of Service Attack (DOS): It is the attack in which the attacker compose some calculating or memory resource much more busy or much full to control legitimate requests, or contradict legitimate users access to a machine.

B. Remote to Local Attack (R2L): It take place only when an attacker who has the endowment to transmit packets to a machine over a network but who does not have an explanation on that machine develop some susceptibility to add local admittance as a user of that machine.

C. Probing Attack: Such type of attack occurs to gather information regarding a network of computers for the noticeable purpose of circumventing its security controls.

D. User to Root Attack (U2R): It is a class of develop in which the attacker starts out with admittance to a common user account on the system and is proficient to widen some susceptibility to gain root access to the system.

TABLE 1
DETAILS OF ATTACKS OF LABELED RECORDS

Category of Attack	Attack Name
Denial of Service(DOS)	Neptune, Smurf, Pod, Teardrop, Land, Back
Probe	Port_sweep, IP_sweep, N_map, Satan
U2R	Buffer overflow, LoadModule, Perl, Rootkit
R2L	Guesspassword,Ftp_write,I_map,Phf,Multi_ho

	p,Warezmaster,Warezclient

5 PROPOSED WORK

Sometimes hidden data occur in classification process of KDD data resulting in generation of false correlations of features so that discovery of intrusion detection process will not be up to the mark. The disadvantage of extra features is that it restrains huge time for the process of computing and it impacts the precisions of IDS. Here feature selection advances the more classification precision by searching for the best features, which best classifies the training data. To overcome the problem of extra features in the proposed system calculate the probability of each and every independent attribute, then entropy has been deliberated and lastly information gain has been calculated for all attributes disjointly. And the logic applied behind this implies that if calculated gain is very less then that type of attribute will not be contributed for the preprocessing of data.

The general equations for entropy and gain are

$$E(T,X) = \sum_{c \in X} P(c)E(c) \quad (1)$$

$$Gain(T,X) = Entropy(T) - Entropy(T,X) \quad (2)$$

5.1 Algorythm Steps

Step 1: Generate reduced dataset 'X1reduced dataset' from a database.

Step 2: Set a learning algorithm to individual pattern for test dataset.

Step 3: Set a learning algorithm to individual pattern training dataset.

svmStruct = svmtrain(X1(train(:,i1),:),groups(train(:,1))

Step 4: Object with unknown found to do with each of the X1 classifiers predictions.

Step 5: Select most repeatedly predicted samples.

5.2 KNNGA Steps

Step 1: Generate reduced dataset 'X1reduced dataset' from a database.

Step 2: Set a learning algorithm to individual pattern for test dataset.

Step 3: Set a learning algorithm to individual pattern training dataset.

svmStruct = svmtrain(X1(train(:,i1),:),groups(train(:,1))

Step 4: Object with unknown found to do with each of the X1 classifiers predictions.

Step 5: Select most repeatedly predicted samples.

KNNGA steps

Step 1: Initialize population = *X1*

Step 2: Apply genetic search into selected dataset

Step 3: Apply KNN classifier for testing of all five data segments which is classified or misclassified data.

Step 4: Organize each attribute according to their ranks.

Step 5: Select higher ranked attributes.

Step 6: Apply KNNGA () on the each five subset of the attributes to enhance the accuracy level.

Step 7: If (knnga_classifier (class_knn) > knn_classifer (class_knn))

```
{
        data_class = class_knn;
        }
    Else {
        data_class = class_knnga;
}
```

Step 8: Perform the reproduction

Step 9: Apply crossover operator

Step 10: Perform mutation then produce new population *X'1*

Step 11: Calculate the local maxima for each category.

Repeat the steps till iteration is not finished

Step 12: For each test *X'1*, start all trained base models then prediction of result by combining of all trained models, and

separate the misclassified by optimized KNNGA.

Classification: Majority occurrences.

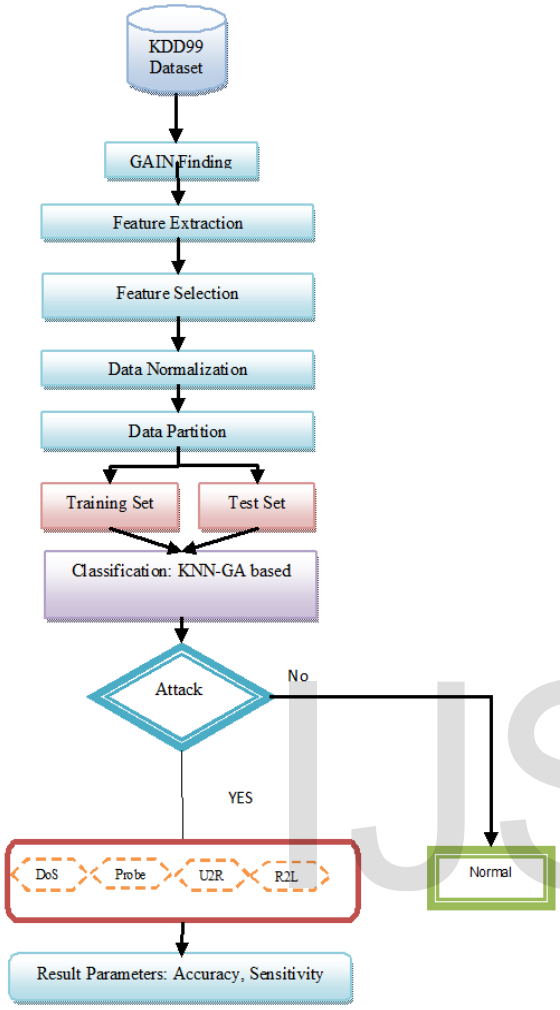

Figure. 5.1 Block diagram

6 Results

To evaluate the performance of the proposed methodology, KDD 99 dataset is used and the simulation of these experiments is implemented on MATLAB2012A, Neural network toolbox in Intel core I5-5200U 2.20 GHz processor having 4 GB of RAM and windows8.1 home basis operating system. The comparison of the proposed approach and existing system is done using performance measuring parameter accuracy which is defined as the percentage of alerts triggered for each different flow cutoff out of the total number of alerts (2252) in the trace.

Here, Figure 6.1 shows that total no. of samples v/s accuracy which shows the comparisons between proposed approach and existing system in which we analyze that our approach is approx 5% more efficient.

TABLE 2
SIMULATION ENVIRONMENT

Sr. No.	Parameter	Values
1	Data size	49119
2	Training ratio	0.5%
3	Features of KDD 99 Dataset	1 to 41
4	Selection ratio	0.5%

TABLE 3
ACCURACY RESULT OF PROPOSED APPROACH

Class	Existing	Proposed
dos	95.77	99.73
probe	95.69	99.7
U2R	95.82	99.72
R2L	95.81	99.72

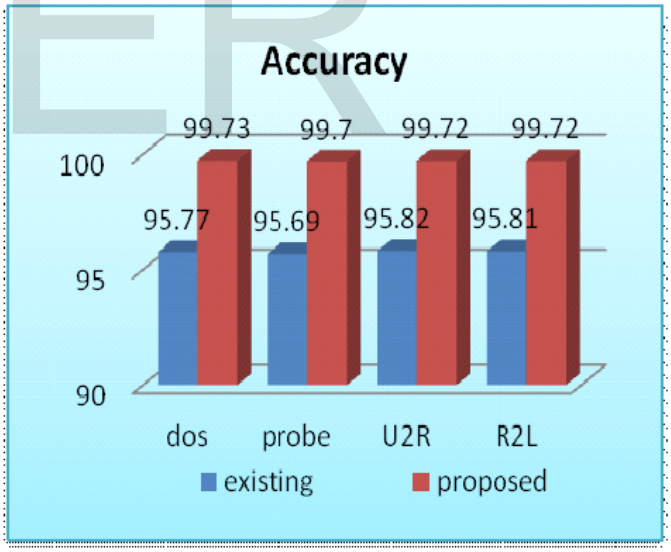

Figure 6.1: Comparison graph b/w propose and existing system

7 Conclusion

Intrusion detection is very crucial part of securing a computer network from thrashing. KDD CUP'99 is most popular dataset since its release till now to test and implement algorithms claiming betterment of intrusion detection process. To overcome the problem of processing overload by using all 41 features of dataset a mathematical solution is proposed. By calculating entropy and gain of every feature, most important fea-

tures are selected. Then KNNGA is applied to train, test and classify the data into various attack types and normal category. Thus the accuracy is improved by about five percent, lessening false alarms rates and improving detection rate is achieved in this way.

REFERENCES

[1] Axelsson, S., "Intrusion Detection Systems:A Taxonomy and Survey," Technical Report No 99-15, Dept. of Computer Engineering, Chalmers University of Technology, Sweden, March 2000.

[2] Lunt, T. F., "Detecting Intruders in Computer Systems," in proceeding of 1993 Conference on Auditing and Computer Technology, 1993.

[3] Sundaram, A. "An Introduction to Intrusion Detection," The ACM Student Magazine, Vol.2, No.4, April 1996. http://www.acm.org/crossroads/xrds2-4/xrds2-4.html.

[4] Porras, P. A., "STAT: A State Transition Analysis Tool for Intrusion Detection," MSc Thesis, Department of Computer Science, University of California Santa Babara, 1992

[5] Dorothy E. Denning, "An Intrusion Detection Model," In IEEE Transactions on Software Engineering, Vol.SE 13, Number 2, page 222-232, February 1987.

[6] Divyatmika, Manasa Sreekesh, "A Two-tier Network based Intrusion Detection System Architecture using Machine Learning Approach", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016 in proceeding of IEEExplore.

[7] Wenke Lee and Salvatore J. Stolfo "Data mining approaches for intrusion detection", In Proceedings of the
7th USENIX Security Symposium - Volume 7, SSYM'98, pages 6–6, Berkeley, CA, USA, 1998.

[8] Rajesh Wankhede, Vikrant Chole, "Intrusion Detection System using Classification Technique", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016.

[9] Mrutyunjaya Panda and Manas Ranjan Patra, "Comparative Study Of Data Mining Algorithms For Network Intrusion Detection" First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008.

[10] G.V. Nadiammai, M. Hemalatha, "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian Informatics Journal 2015, in proceeding Elsevier Pp 37–50.

[11] Mouhammd Alkasassbeh, Ahmad B.A Hassanat, Ghazi Al-Naymat, Mohammad Almseidin, " Detecting Distributed Denial of Service Attacks Using Data Mining Techniques", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, 2016.

[12] M. R. Norouzian and S. Merati, "Classifying attacks in a network intrusion detection system based on artificial neural networks," in Advanced Communication Technology (ICACT), 2011 13th International Conference on, pp. 868–873, IEEE, 2011.

[13] Jianfeng Pu, Lizhi Xiao, Yanzhi Li and Xingwen Dong "A Detection Method of Network Intrusion Based on
SVM and Ant Colony Algorithm", National Conference on Information Technology and Computer Science (CITCS 2012) Published by Atlantis Press.

[14] Ayman I. Madbouly, Amr M. Gody, Tamer M. Barakat, "Relevant Feature Selection Model Using Data
Mining for Intrusion Detection System", International Journal of Engineering Trends and Technology (IJETT) – Volume 9 Number 10 - Mar 2014.

[15] S. Selvakani Kandeeban, Dr. R.S. Rajesh : "a genetic algorithm based elucidation for improving intrusion detection through condensed feature set by KDD99 dataset", information and knowledge management ISSN 2224-5758, ISSN 2224-896X Vol. 1, No.1, 2011, www.iiste.org.

[16] Mouaad KEZIH, Mahmoud TAIBI "evaluation effectiveness of intrusion detection system with reduced dimension using data mining classification tools", 2nd International Conference on Systems and Computer Science (ICSCS) Villeneuve d'Ascq, France, August 26-27, 2013; 978-1-4799-2022.