Anomaly based Intrusion Detection System using Genetic Algorithm and K-Centroid Clustering

Biswapriyo Chakrabarty
Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet

Omit Chanda
Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet

Md. Saiful Islam
Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet

ABSTRACT
Internet is being expanded because of the enhancement of today's networks and with these expansion different types of unauthorized activities building up to make the network vulnerable. Many researchers are working around the world to protect the systems from any kind of unauthorized access. In this study we have implemented an Intrusion Detection System based on K-Centroid Clustering and Genetic Algorithm to achieve a better detection rate and false positive rate. In our system training set is classified into different clusters based on K-Centroid clustering and then GA is performed to check each connection of the test set and finally result has been obtained for every specific connection. We have used both Kdd99Cup and NSLKDD dataset to get the experiment result of our system. Finally analyzing with those data we have got a decent detection rate in our implemented system.

General Terms
Network Security

Keywords
Computer Security, Intrusion Detection, Intrusion Detection Systems, Genetic Algorithm, K-centroid Clustering

1. INTRODUCTION
Computer networks expanding at a prodigious rate and with this expansion different types of users are being attached with this network. For the congeniality of the mountainous number of users a strict security policy should be maintained. Thinking about the safety of these thousands of users, several security techniques brought to us in some respects but it isn't seemed to be sufficient. To provide the preservation of the network systems intrusion detection has become a significant technique and gradually its taking the utmost space to make the systems secured. Intrusion detection is a technique used to monitor network traffic, identifying unauthorized access and different malicious behaviors make the networks intrusion free[1]. Intrusion detection system (IDS) is that using which the technique is performed. IDS must differentiate between authorized and unauthorized activities to make the system accurate[2]. Without distinguishing both malicious and non-malicious packets from the network traffic IDS will not be irreproachable.

There are two categories of intrusion detection systems: misuse detection and anomaly detection. Misuse detection techniques detect intrusions using known patterns and on the other hand anomaly detection technique identifies a connection as an attack if it will strongly be deviated from the normal data. IDSs can also be divided into two categories. Host based IDS detects intrusions from any particular computer and Network based IDS checks intrusion from network traffic[3].

Several soft computing techniques can be used to implement intrusion detection systems. Different kinds of soft computing techniques such as Genetic Algorithm (GA), Artificial Neural Networks (ANN), Fuzzy Logic etc. used in implementing different intrusion detection approaches. In this work we have developed a system based on GA and another machine learning technique K-Centroid clustering. There are many features used in GA to make it very compatible for intrusion detection purpose. Like robustness to noise, self-learning capabilities, and the fact that initial rules can be built randomly so there is no need of knowing the exact way of attack machinery at the beginning[1].

In intrusion detection, Genetic Algorithm takes a weighty space to ameliorate the vogue of this field. It operates on a large number of populations and find the best fitted one by performing different genetic process such as crossover, mutation, selection etc.[5],[6]. A lot of generations can be evolved to perform the algorithm and in each generation better adapted populations selected from previous generations and process performed again and again until there only the best individual is available.

There are different data-sets Kdd99Cup and NSLKDD used in different implementation to evolve new rules. Kdd99Cup is the dataset with many handicaps such as worthless features, some missing attacks in train set, many redundant records etc. causes complication to determine rules for IDS[14]. On the other hand, NSLKDD dataset suggested solving those problems of the KDD'99 data set and it can be applied as an effective dataset to create a clear comparison with intrusions and normal data in developing intrusion detection systems. Moreover, the number of test and train data in NSLKDD is reasonable for the system implementation[13]. In spite of several disadvantages in Kdd99Cup, we have used it to get better comparison with many existing system as many of the present systems had used Kdd99Cup to show their experimental result. We have also used latest NSLKDD dataset in order to experiment our system because using a better faultless dataset the actual detection rate of a system can be measured.

2. RELATED WORKS
Many researchers are working in this field to improve the security of Networks and increase the accuracy rate towards the finest one day by day. Several important works have been studied for our research introducing below.

Ren Hui Gong et al.[3] used the support-confidence framework as fitness function in their GA based intrusion detection system. The results of the system were calculated based on 1998 DARPA data sets. The speedy adaptation with the complicated and rapidly changing intrusion types makes the proposing system advantageous. But the redundancy of connections in training dataset can make the generated rules biased.

M A. Chittur et al.[4] implemented an anomaly detection approach based on GA. In that study, the system was run on 1999 Knowledge Discovery in Database (KDD) Cup data. To check the malicious behavior of a connection a threshold value must be established and if the fitness value of any connection deviated from the threshold value figured as a malicious attack. The system must be known about the threshold value because it is the only feature to detect malicious behavior of a connection but the determining process of threshold value is not so easy.

Li et al.[5] described a method using GA to detect anomalous network intrusion[3][15]. The approach includes both quantitative and categorical features of network data for generating classification rules. During the encoding the system takes both temporal and spatial information of connections that's why the proposed algorithm has become more helpful to identify the malicious behavior of different network connections.

Lu and Traore[6] used historical network dataset using GP to derive a set of classification[?]. They used support-confidence framework as the fitness function and accurately classified several network intrusions. But their use of genetic programming made the implementation procedure very difficult and also for training procedure more data and time is required.

Jonatan Gomez and Dipankar Dasgupta[7] used fuzzy logic and fuzzy rules in their system. The main problem with this approach is to generate good fuzzy classifiers to detect intrusions. This paper proposed a technique to generate fuzzy classifiers using genetic algorithms that can detect anomalies and some specific intrusions.

Mohammad Sazzadul Hoque et al.[16] used evolution theory in order to filter the traffic data and thus reduce the complexity. To implement and measure the performance of the system they used the KDD99 benchmark dataset and obtained reasonable detection rate. But KDD99Cup data could not be a standard dataset to get a satisfactory accuracy rate because of the redundancy issue in the dataset.

Wang et al.[10] proposed a system based on Artificial Neural Network and fuzzy clustering. They have used KDDCup99 dataset for their experimental purpose. Because of using the problematic KDDCup99 dataset the rate of accuracy may be a little bit of confusing.

3. WORKING APPROACHES

Several machine-learning paradigms including clustering, neural networks, linear genetic programming (LGP), support vector machines (SVM), Bayesian networks, multivariate adaptive regression splines(MARS), fuzzy inference systems (FISs) etc. have been investigated for the design of IDS. In this study, we have worked with Clustering and Genetic Algorithm for the purpose of Intrusion Detection System. Moreover we have used both KDDCup99 and NSLKDD dataset to test our proposed system. Different key concepts of our system are describing below.

3.1 Clustering

Clustering is used to build different groups for a specific normal or intrusive data. There are various types of patterns can be existed for any particular attack type or normal connection and because of the large dissimilarity of those patterns each analogous type of connection isn't included into a single cluster. So accomplishing Clustering different clusters constructed for a particular similar type of connection. Then GA performed on those clusters to identify a specific target class for a newly occurred connection.

3.2 Genetic Algorithm

Genetic algorithm[3][4],[9] takes its concept from bioinformatics to evolve a population having some initial individuals where every individual treated to be a chromosome and when different types of genetic operators engaged to perform some tasks such as mutation, crossover, selection etc. then a high quality individual can be picked out.

To start this algorithm firstly a population of randomly generated individuals must be initialized. After initialization evolution started for each generation. During every generation some basic operations established to each chromosome with certain probabilities. Firstly, selection performed to select some best fitted individuals based on a fitness function. Secondly, selecting a number of an individual another operation crossover has to be performed and that's why individuals paired with each other and new individuals produced by exchanging their genes around one or more randomly selected crossing point. Lastly a number of individuals selected and mutation performed into some random position of a particular individual.

There are three factors have to be used efficiently for the success of the algorithm. Those are: 1) Selection of fitness function 2) How individuals represented and 3) What will be the values of genetic parameters. All these factors vary based on different systems. In our implementation standard deviation used as fitness function. Individuals also called chromosomes accumulated with a number of genes, here different features of dataset involved as genes. Using a large number of experiments mutation rate, crossover rate selected to increase the accuracy rate of the implementation.

3.3. Data Set

In this implementation two dataset KDDCUP99 and NSLKDD used to get experimental results. KDD CUP 1999[11],[12] dataset holds about five million connection records as training data and two million as test data. In each connection there exists 41 numbers of features and a label is used to specify the target class of a connection. Label of any specific connection can be either normal or any attack type. There are four categories of attacks in the dataset: 1) Denial of Service (DoS): Its created by making any memory resources busy to acknowledge allowable users. 2) Probe: Searching vulnerabilities of any systems knowing host and port information. 3) Remote to Local (R2L): Unauthorized access from any remote systems in order to utilize system's vulnerabilities. 4) User to Root (U2R): unauthorized access to local super user (root) privileges using system's sensitivity.

There are many disadvantages in KDD CUP 1999 like worth-less features, missing attacks in train set, many redundant records etc.[14] create many problems to generate IDS rules and NSLKDD[13] dataset suggested to solve those inherent problems of the KDD'99 data set. It is the latest version of the dataset. It consists of the same features as KDD 99[11]. NSL KDD dataset also includes 41 features and one target class in

its each connection. It doesn't include redundant records in the train set so the classifiers will not be biased towards more frequent records. The number of records in both the train and test set are reasonable which makes the systems affordable to run with the whole dataset.

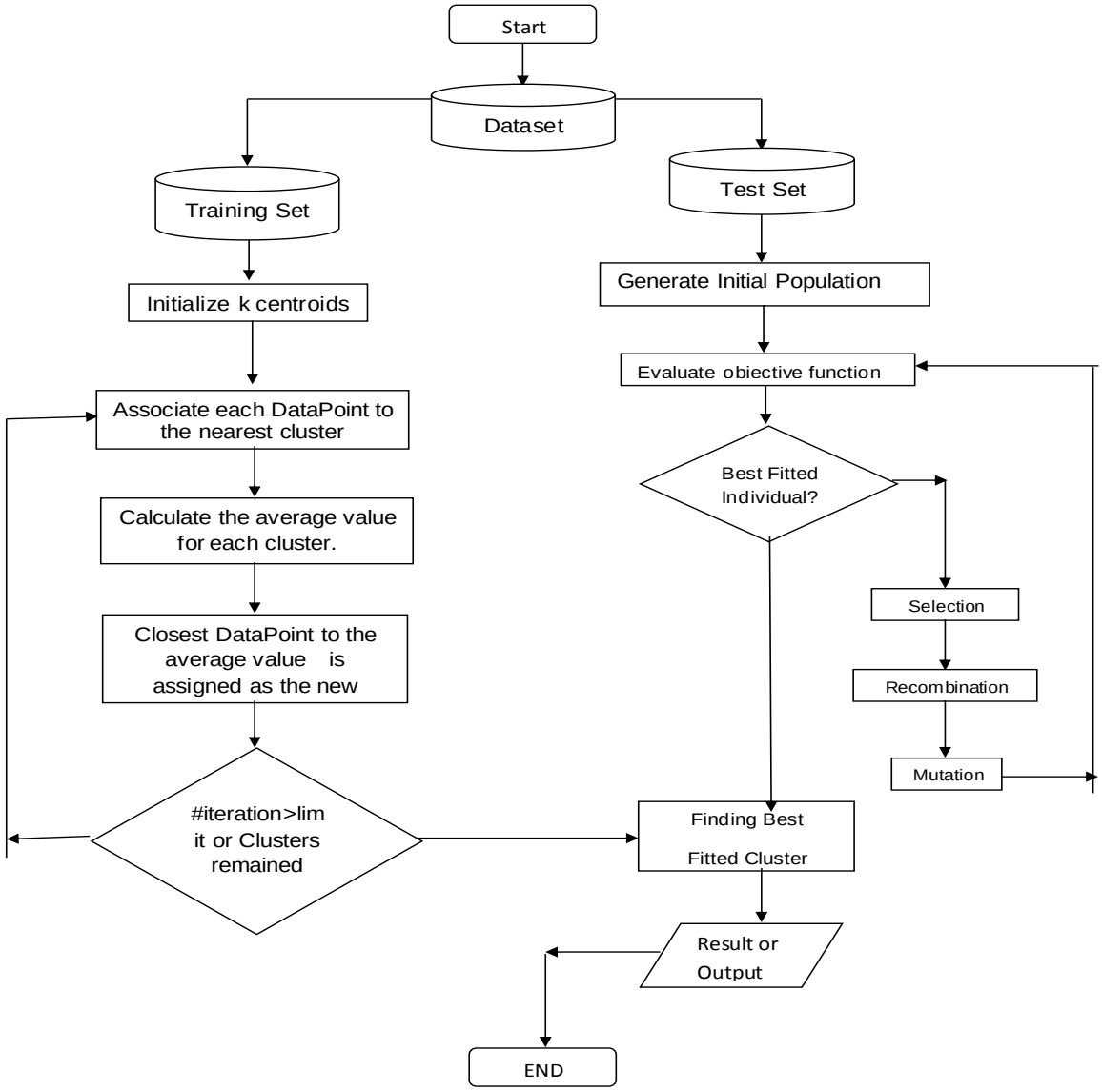

Fig 1: Flowchart of our IDS

4. FRAMEWORK OF OUR IDS

Here we will discuss about the overview of our system. Firstly our framework will be elaborated which consists of three modules such as K-Centroid clustering module, GA module and Target finding module.

4.1. Framework of IDS based on K-Centroid clustering and GA

In our system Clustering makes a certain number of groups to determine some rules those are used to check whether a connection malicious or normal. To perpetrate this task a train set must be needed. Using those training data our system will be trained and a specific number of groups will be ready for future works. Committing clusters GA performed to fulfill the next steps and resolves a connection whether it is normal or intruder. The architecture of our system is shown in following figure Framework of our IDS crosses 3 stages in detecting malicious behavior. Those steps are included below:

Step 1: From a training dataset extracting data and integrating K-Centroid clustering a certain number of clusters constructed in K-Centroid Clustering module.

Step 2: Here a connection is picked out from the test set and a number of individuals have been ready to perform different genetic operations and evolved during each generation to get a best fitted individual. When best fitted individual found this step will be ended.

Step 3: In final module a single individual will be compared with different clusters and final result will be obtained for every connection that the specific connection is related with which group.

4.2 Algorithm Overview of the System

Initially a large number of connections initialized from training dataset (line-1). Then a specific number of clusters loaded with several chromosomes. Line 4 to 12 performed K-Centroid clustering technique to finalize all clusters for future work. After establishing all clusters testing process of the system started from line number 13. In testing procedure a connection retrieved from test dataset and an initial population generated for performing Genetic Algorithm to find the ultimate attack type of a connection. In line 15-24 different genetic operators' selection, mutation, crossover etc. operated to get the best fitted chromosome from the population. After getting the best fitted one finally line 25-31 performed to find the type of the testing connection. If it reaches the cluster with normal pattern then it considered as normal otherwise the connection treated as malicious.

1) Algorithm: Intrusion Detection using Clustering & GA

2) Input: KDDCup99 and NSLKDD dataset

3) Output: Whether Connections are Normal or Malicious

```
1: N ← Numberofconnectionsorchromosomesintrainingdataset
2: K ← Numberofclusters
3: Initialize the clusters with certain number of Connections
4: for certain number of times do
5:   for k := 1 to K do
6:     Select record from N
7:     if the record matches with the cluster then
8:       Record will be entered into the cluster
9:     end if
10:  end for
11:  Centroid of all clusters changed
12: end for
13: R ← recordtobetested
14: Generate M chromosomes using R
15: while M ≥ 1 do
16:   T ← (M + 1)/3
17:   t ← T
18:   while t ≤ 2 do
19:     Select two chromosomes from t
20:     Apply crossover to them
21:     Apply mutation to the chromosomes
22:     t ← t − 2
23:   end while
24:   M ← T
25: end while
26: C ← bestfittedchromosomeproducedfrompreviousoperations
27: T ← 1 {C has become a member of 1st cluster}
28: for k := 2 to K do
29:   if C more similar with the cluster k then
30:     T ← k
31:   end if
32: end for
33: if T is the cluster of malicious connections then
34:   T connection is intrusive
35: else
36:   T connection is normal
37: end if
```

5. EXPERIMENTAL RESULTS

System which was experimented on two popular dataset KDD Firstly we have used KddCup99 dataset in our experiment to train and test the system. The whole number connections of the dataset tested by making 100 and 200 number of clusters in two times. In that time we have counted only normal and malicious connection to calculate the total accuracy rate of our system.

Testing the system based on KDDCup99 dataset we have got more than 90% overall accuracy rate in our system. False positive rate was also negligible.

Table 1. Anomaly Detection and Accuracy Rate Analysis for KDDCUP99

Connection		No. of Clusters-100		No. of Clusters-200	
Type	**Number**	**Detected**	**Accuracy**	**Detected**	**Accuracy**
Normal	9711	8291	85.30%	8270	85.10%
DOS	5734	5212	91%	5290	92.30%
Probe	1106	984	89%	1060	95.80%

After testing the system using KDDCup99 dataset we have taken NSLKDD dataset to experiment our system more fruitfully. Comparatively In that case we have got a decent accuracy rate by making 100, 150 and 200 number of clusters in three times. Using 100 clusters we got approximately 85% accuracy rate, on the other hand using 150 and 200 clusters we got approximately 86% accuracy rate

Table 2. Anomaly Detection and Accuracy Rate Analysis for NSLKDD

Connection		No. of Clusters-100		No. of Clusters-200	
Type	**Number**	**Detected**	**Accuracy(%)**	**Detected**	**Accuracy(%)**
Normal	9711	8291	85.30%	8270	85.10%
DOS	5734	5212	91%	5290	92.30%
Probe	1106	984	89%	1060	95.80%
R2L	2536	1682	66.30%	1941	76.50%
U2R	54	32	60%	16	31.50%
Total	19141	16201	84.50%	16577	86.60%

6. CONCLUSION

In this research we have implemented an intrusion detection system which was experimented on two popular dataset KD-DCup99 and NSLKDD. We have proposed the hybrid system based on two machine learning modules K-Centroid clustering and Genetic Algorithm. In KDDCup99 dataset the experiment result may be biased because of the redundancy rate of records into the dataset that's why we have also tested our system using one of the latest datasets NSLKDD. Using both datasets we have got a magnificent accuracy rate compared to the other existing systems. Some other machine learning techniques are also becoming popular for the improvement of network security systems and in future we can also study with those important fields so that we can get an intrusion detection system with better accuracy rate and better false positive rate.

7. REFERENCES

[1] Zorana Bankovic, Dusan Stepanovic, Slobodan Bojanic, Octavio Nieto-Taladriz. "Improving network security using genetic algorithm approach", Computers and Electrical Engineering, Pg 438-451, July 2007.

[2] S.Owais, V.Snasel, P.Kromer and A. Abraham, "Survey Using Genetic Algorithm Approach in Intrusion Detection Systems Techniques", 7 th Computer Information Systems and Industrial Management Applications, 2008, IEEE , pp.300-307.

[3] A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection Ren Hui Gong, Mohammad Zulkernine, Purang Abolmaesumi

[4] Chittur A. Model Generation for an Intrusion Detection System Using Genetic Algorithms, http://www1.cs.columbia.edu/ids/publications/gaids-thesis01.pdf, accessed in 2006.

[5] W. Li, "Using Genetic Algorithm for Network Intrusion Detection". "A Genetic Algorithm Approach to Network Intrusion Detection". SANS Institute, USA, 2004.

[6] W. Lu, I. Traore, "Detecting New Forms of Network Intrusion Using Ge-netic Programming". Computational Intelligence, vol. 20, pp. 3, Blackwell Publishing, Malden, pp. 475-494, 2004.

[7] Jonatan Gomez and Dipankar Dasgupta. Evolving fuzzy classifiers for intrusion detection. In Proceedings of the 2002 IEEE Workshop on Information Assurance, West Point, NY, USA, 2002.

[8] S. Selvakani Kandeeban & R. S. Rajesh, "A Mutual Construction for IDS Using GA", 2011

[9] M. Crosbie and E. Spafford, "Applying Genetic Programming to Intrusion Detection", Proceedings of the AAAI Fall Symposium, 1995

[10] Gang Wang, Jinxing Hao, Jian Ma, Lihua Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy cluster-ing".

[11] KDD cup 1999 data http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[12] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[13] http://nsl.cs.unb.ca/NSL-KDD/

[14] A COMPARISON STUDY FOR INTRUSION DATABASE (KDD99, NSL-KDD) BASED ON SELF ORGANIZATION MAP (SOM) ARTIFICIAL NEURAL NETWORK

[15] B. Abdullah, I. Abd-alghafar, Gouda I. Salama, A. Abd-alhafez, "Per-formance Evaluation of a Genetic Algorithm Based Approach to Network Intrusion Detection System", 2009'

[16] Mohammad Sazzadul Hoque, Md. Abdul Mukit and Md. Abu Naser Bikas , "AN IMPLEMENTATION OF INTRUSION DETECTION SYS-TEM USING GENETIC ALGORITHM" , Dept. of Computer Science & Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh.