International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Hybrid System Using Genetic Algorithm for Anomaly Intrusion Detection

Arpitha J[1], Nagaraj Naik [2]

[1]M.Tech Student, Dept. of Computer Science & Engineering, Mangalore Institute of Technology and Engineering Mangalore, Karnataka, India

[2]Senior Assistant Professor, Dept. of Computer Science & Engineering, Mangalore Institute of Technology and Engineering Mangalore, Karnataka, India

Abstract: ***This paper proposes a new approach to design the system using a hybrid of misuse and anomaly detection for training of normal and attack packets respectively. The hybrid intrusion detection system combines the K-means and the genetic algorithm for anomaly detection. This algorithm operates on the KDD-99 Data set; this data set is used worldwide for evaluating the performance of different intrusion detection systems. The system can detect the intrusions and further classify them into four categories: Denial of Service (DoS), U2R (User to Root), R2L (Remote to Local), and probe. The main goal is to reduce the false alarm rate of IDS.***

Keywords: Clustering, Classification, K-Means, Genetic Algorithm, Detection rate, False alarm rate, Intrusion detection, data mining, KDD cup 99.

1. Introduction

The internet has become a part of daily life and an essential tool today. As more and more sensitive data continues to get stored and manipulated online; the need for an increase in security of network systems is getting more and more importance day by day. The following 3 functionalities must be provided essentially by any secure network [1].

- ***Data confidentiality:*** Data access must strictly be done by authorized users. Eavesdroppers and intruders must mot gain any important information.
- ***Data availability:*** Authorized system users must be able to access and use any resource at any point in time.
- ***Data integrity:*** Corruption and data loss of information must be prevented. Exactness of data must be preserved.

2. Intrusion Detection System

An Intrusion Detection System (IDS) is a defence system which inspects the activities in a system for suspicious behavior or patterns that may indicate system attack or misuse and then notify intrusion prevention system (IPS) or network security administrator so that suitable actions can be taken against the attacks. Following are the 2 important approaches to detect intrusions [4].

A. Misuse detection

In Misuse detection patterns for different malicious behaviours are built first, and then the attacks are detected based on these predefined patterns. Misuse detection is very effective in avoiding an immense amount of false alarms and provides a great accuracy. However, misuse detectors can only detect attacks whose signatures are known. Any variations of the common attacks go undetected. Signatures of new attacks must be constantly updated.

B. Anomaly Detection

In anomaly detection, a normal profile which describes the behaviour of the system under normal conditions is constructed in advance. Any significant aberrations from such expected behaviour are reported as possible attacks. The major advantage of this approach is that with fewer details unusual behavior can be easily detected, thereby effectively reducing the storage and maintenance cost. But it requires a large amount of "training sets" to effectively characterize normal behavior. Another shortcoming of anomaly detection is its high false alarm rate.

The main purpose of intrusion detection is to detect future attacks which have led to learning techniques. The intrusion detection model cannot adapt to the network behavior pattern. So in order to detect new attacks and continually adapt with the new network behavior, we propose a hybrid intrusion detection system that is composed of misuse and anomaly detection system. This system combines the merits of misuse and anomaly detection. Our goal is not only to obtain high detection rate (DR) on malicious activities but also to reduce the False Positive Rate (FPR) on normal computer usage from network traffic.

3. Related Works

Hybrid intrusion detection systems comprise of misuse detection and anomaly detection systems that can detect both known and unknown intrusions.

Most common example is a combination of Naïve Bayes Classifier and K-Means [6]. Here after grouping the data into suitable clusters, classifier is applied for classification purpose.

Tsai and Lin employ K-Means clustering to cluster data instances into k-clusters [7]. Next, the research trains the new dataset, which consists of only the centers of cluster with Support Vector Machine (SVM). They managed to obtain high accuracy rate for almost to all attack types. This approach offers high detection rate but comes with high false alarm rate.

Intrusion detection based on Fuzzy SVMs (FSVM) was proposed by Shaohua et al. [8] to improve the classification accuracy. The purpose of the clustering algorithm is to construct a new training set using centers of clusters. This new set will then be trained with FSVM to obtain a support vector. Although their results have proved that this method has increased the accuracy rate, it is not of an acceptable percentage.

Various techniques have been proposed in the intrusion detection field and related work; but there are still room to improve the accuracy and detection rate as well as the false alarm rate.

4. Hybrid Learning Approach

Anomaly learning approaches are able to detect attacks with high accuracy and high detection rates. However, the rate of false alarms is also high. In order to maintain the high accuracy and detection rate while at the same time reduce the false alarm rate, we propose a combination of two techniques.

For the first stage in the proposed hybrid learning approach, we grouped similar data instances based on their behaviors by utilizing a K-Means clustering as a pre-classification component. Next, using genetic algorithm we classify the resulting clusters into attack classes as a final classification task. We found that data which have been misclassified during
the earlier stage may be correctly classified in the subsequent classification stage.

A. K-Means Clustering

K-Means is one of the simplest unsupervised learning methods among all partitioning based clustering methods. It classifies a given set of *n* data objects in *k* clusters, where *k* is the number of desired clusters and it is required in advance. A centroid is defined for each cluster. All the data objects are placed in a cluster having centroid nearest (or most similar) to that data object. After processing all data objects, k-means, or centroids, are recalculated, and the entire process is repeated. All data objects are bound to the clusters based on the new centroids. In each iteration centroids change their location step by step. This process is continued until no any centroid move. As a result, *k* clusters are found representing a set of *n* data objects. An algorithm for k-means method is given below

Algorithm:
Input: '*k*', the number of clusters to be partitioned; '*n*', the number of objects.
Output*:* A set of '*k*' clusters based on given similarity function.
Steps:
i) Arbitrarily choose '*k*' objects as the initial cluster centers;
ii) Repeat.
a. (Re) assign each object to the cluster to which the object is the most similar; based on the given similarity function.
b. Update the centroid (cluster means), i.e., calculate the mean value of the objects for each cluster.
iii) Until no change.

B. Genetic Algorithm

Genetic Algorithms (GAs) are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. The basic concept of GAs is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the fittest. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem.

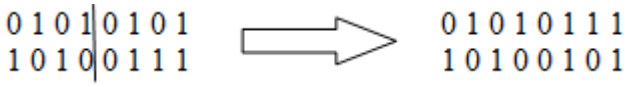
GAs are one of the best ways to solve a problem for which little is known. They are a very general algorithm and so will work well in any search space. GA is stochastic search algorithm modeled on the process of natural selection, which underlines biological evolution [9]. GA has been successfully applied in many research, optimization and machine learning problems. GA works in an iterative manner by generating new populations of strings from old ones.

Standard GA apply genetic operators such selection, crossover
and mutation on an initially random population in order to compute a whole generation of new strings. GA runs to generate solutions for successive generations. The probability of an individual reproducing is proportional to the goodness of the solution it represents. Hence the quality of the solutions in
successive generations improves. The process is terminated when an acceptable or optimum solution is found.

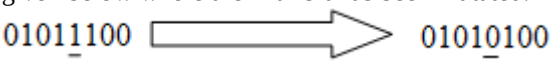
GA is appropriate for problems which require optimization, with respect to some computable criterion. The functions of genetic operators are as follows:

1) *Selection:* Selection deals with the probabilistic survival of the fittest, in that, more fit chromosomes are chosen to survive. Where fitness is a comparable measure of how well a chromosome solves the problem at hand.
2) *Crossover:* This operation is performed by selecting a random gene along the length of the chromosomes and swapping all the genes after that point.
3) *Mutation:* Alters the new solutions so as to add stochasticity in the search for better solutions. This is the chance that a bit within a chromosome will be flipped (0 becomes 1, 1 becomes 0). Genetic algorithms are a method of "breeding" computer programs and solutions to optimization or search problems by means of simulated evolution. Processes loosely based on natural selection, crossover, and mutation are repeatedly applied to a population of binary strings which represent potential solutions.

In standard genetic algorithm two parents are selected at a time and are used to create two new children to take part in the next generation. The offsprings are subjected to the operator crossover with a pre-specified probability of crossover. Single point crossover is the most common form of this operator. It marks a random crossover spot within the size of chromosome and exchanges the bits on the right of the spot as shown below



Mutation operator is applied to all the children after crossover. It flips each bit in the individual with a pre-specified probability of mutation. An example of mutation is given below where the fifth bit has been mutated.



The procedure is repeated till number of individuals in the population is completed. It finishes one generation in the genetic algorithm. GA is run till a stopping criterion is satisfied that may be defined in many ways.

5. Experiments & Results

A. Dataset Description

In our experiments, the KDD Cup'99 benchmark dataset KDD [10] is chosen for evaluation and comparison between the proposed approaches and the previous approaches. The entire KDD data set contains an approximately 500,000 instances with 41 features. The training dataset contains 38 types of attacks, while the testing data contains more than 28 types of additional attack. Further description for the available features and intrusion instances can be found in [10].

In order to demonstrate the abilities to detect different kinds of intrusions, the training and testing data covered all classes of intrusion categories as adopted from [11] as follows:

DoS Attack*:* Denial-of-service attack is a type of attack on a network that is designed to bring the network to its knees by flooding it with useless traffic. Usual target for such kinds of attacks are high profile web servers such as banks. Though DoS attacks do not typically result in the theft or loss of information they can cost the victim a great deal of time and money.

Remote to User (R2L): Here an attacker tries to gain access to the local machine from a remote machine by some unauthorized means. Social engineering is one such attack.

Probes**:** It is a class of attacks where an attacker continuously scrutinizes a network until he finds all the vulnerabilities present. Attacks are then staged by exploiting these loopholes.

User to Root (U2R): Here an attacker tries to attack where a local user on a machine is able to obtain privileges normally reserved for the UNIX super user or the Windows NT administrator.

6. Results and Analysis

In order to evaluate the performance of this method we have used KDD99 data set. First we apply the K-means clustering algorithm on the features selected. After that, we will generate the new records and then we classify the obtained data into Normal or Anomalous clusters by using the Hybrid classifier.

An efficient IDS requires high accuracy and detection rate as well as low false alarm rate. In general, the performance of IDS is evaluated in terms of accuracy, detection rate, and false alarm rate as in the following formula:

Accuracy = (TP+TN) / (TP+TN+FP+FN)
Detection Rate = (TP) / (TP+FP)
False Alarm = (FP) / (FP+TN)

Table 1 show the categories of data behavior in intrusion detection for binary category classes (normal and attacks) in terms of true negative, true positive, false positive and false negative.

Table 1: General Behavior of Intrusion Detection Data

Actual	Predicted Normal	Predicted Attack
Normal	TN	FP
Intrusions(attacks)	FN	TP

- True positive (TP) when attack data is detected as attack
- True negative (TN) when normal data is detected as normal
- False positive (FP) when normal data is detected as attack
- False negative (FN) when attack data is detected as normal

Table 2: Result For K-means+ Genetic Data Classifier Using Normal and Attack Class

No of original records	No of new generated records	True Positive	True negative	Outliers
21120	30000	28850	548	548
42273	50000	47395	350	2255
126819	140000	96962	1282	41756
380461	200000	361953	2485	16023

Table 3: DR, FAR and Accuracy

	Detection Rate	False Alarm Rate	Accuracy %
1	0.039046586	0.32273263	98.58
2	0.045417905	0.13435704	98.42
3	0.3010136	0.0297876	89.74
4	0.0423916	0.13426626	98.60

The table 1 gives the general behavior of the intrusion detection data. The table 2 gives the results for the k-means and genetic data classifier and also the results of the misclassified (outliers) records. The detection rate, false positive rate, accuracy are calculated from the confusion matrix table using the given formula and results are given in table 3. We can see that there is a sharp increase in detection rate accuracy and decrease in false alarm rate. This shows that our proposed approach is better than the conventional k-Means algorithm.

7. Conclusion

In this paper, we have proposed a hybrid intrusion detection system that combines the merits of anomaly and misuse detection. Anomaly detection have very high false alarm rate. In order to reduce it we have applied the k- Means algorithm for clustering followed by a hybrid classifier, combining genetic algorithm classifier for detecting intrusions. The disadvantage of the existing methods is that the data set in real life has very little difference between

normal and anomalous data. The differences are sometimes so small that the classification algorithms misclassify them and some records are misclassified. We have overcome this problem by classifying in the misclassified records (outliers).

8. Acknowledgment

I am very thankful to my guide Mr. Nagaraj Naik, Sr. Assistant Professor, Department of Computer Science and Engineering, MITE for his cordial support, valuable information and guidance, to prepare this paper and also thankful to Prof. Dr. Nagesh H R, Head of the Department, Computer Science and Engineering, for his valuable and constructive suggestions during the planning and development of this work.

References

[1] Reema Patel, Amit Thakkar, Amit Ganatra," A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.
[2] Manasi Gyanchandani, J.L.Rana, R.N.Yadav," Taxonomy of Anomaly Based Intrusion Detection System: A Review", International Journal of Scientific and Research Publications ISSN 2250-3153, Volume 2, Issue 12, December 2012.
[3] RavindraThool, Kapil Wankhade ,Sadia Patka," An Overview of Intrusion Detection Based on Data Mining Techniques", International Conference on Communication Systems and Network Technologies, 2013.
[4] Poonam Dabas, Rashmi Chaudhary, "Survey of Network Intrusion Detection Using K-Mean Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.
[5] Amanpreet Chauhan, Gaurav Mishra, Gulshan Kumar," Survey on Data Mining Techniques in Intrusion Detection", International Journal of Scientific & Engineering Research Volume 2, Issue 7, July-2011.
[6] Z. Muda, W. Yassin, M.N. Sulaiman, N. I Udzir,"A K-Means and Naïve Bayes Approach for Better Intrusion Detection", Information Technology Journal, 648-655, 2011.
[7] Tsai, C.F. and Lin, C.Y, 2010. A triangle area-based nearest neighbors approach to intrusion detection. *Pattern Recognition*, 43(1): p.222-229.
[8] Shaohua, T., Hongle, D., Naiqi, W., Wei, Z., and Jiangyi, S., 2010. A Cooperative Network Intrusion Detection Based on Fuzzy SVMs. *Journal of Networks*, 5: p.475–483.
[9] Stuart J. Russel, Peter Novig (2008) Artificial Intellegence: A Modern Approach.
[10] KDD (1999). < http://kdd.ics.uci.edu/databases/ -kddcup99/kddcup99.html> [Accessed 5 Jan 2011].
[11] Breiman, L. Et al., 1984. *Classification and regression trees*. Monterey, CA: Wadsworth & Books/Cole Advanced Boks & Software.

Author Profile

Miss Arpitha J completed the Bachelor's Degree in Computer Science & Engineering from Visvesvaraya technological University (VTU). Currently pursuing M.Tech degree in Computer Science & Engineering at Mangalore Institute of Technology, Mangalore under VTU, Belagavi.

Mr. Nagaraj Naik senior assistant professor MITE, Mangalore. Completed his M.Tech in computer science and engineering having 8.5 years of academic experience and his areas of interest are java programming, operating system.