International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 4, July - August 2015 **ISSN 2278-6856**

OUTLIER DETECTION AND SYSTEM ANALYSIS USING MINING TECHNIQUE OVER KDD

manoj mishra , nitesh gupta

NRI collage bhopal

ABSTRACT

The intrusion detection system has been implemented using various data mining techniques which help user to identify or classify various attacks or number of intrusion in a network. KDD dataset is one of the popular dataset to test classification technique s. In this paper our work is done on analysis of different techniques which were used in order to detect outlier and get the IDS system improved based on the system. The paper investigate the algorithm such as – Naïve Baise, FPOP, SCF, K-Mean and A hybrid approach of Genetic and SVM which is combined and proposed by us to find better outlier as compare to other proposed algorithm by different authors. Our work contribute the investigation and analysis of different outlier detection technique over KDD dataset.

Keywords- IDS, KDD dataset, Outlier detection, classification.

I. INTRODUCTION

From security point of view, the increasing demand of network connectivity makes the system insecure. So we need a complement that can cope / prevent the security breaches in system. Unfortunately, in many environments, it may not be feasible to render the computer system immune to all type of intrusions. The motivation behind this project is to develop a complement system i.e. Intrusion detection system (IDS) that can prevent all possible breaks-ins.Generally, Intrusion detection system (IDS) has three component that is detection component, investigating component, and post-mortem component. The detection component identifies security breaches. The investigation component determines exactly what happened based on data from the detection component. This component may also include the gathering of further data in order to identify the security violator. Finally, the post-mortem component analyzes how to prevent similar intrusions in the future. With the emergence and the proven utility of the intrusion detection concept, the detection component is beginning to receive more attention. There are many approaches to find intrusions under detection component. Since volume of data dealing with network is so large, this collection concentrates heavily on the use of data mining in the area of intrusion detection. Classification is one of the effective techniques under data mining that can be used for intrusion detection. Two common knowledge representations for classification technique are IF-THEN rules and Decision tree.

II. INTRUSION DETECTION SYSTEM

The number of hacking and intrusions incidents is increasing year on year as technology rolls out. Maintaining a high level security to ensure safe and trusted communication of information between various organizations becomes a major issue. So Intrusion detection system (IDS) has become a needful component in terms of computer and network security . This report proposed a framework for intrusion detection system that uses decision tree for its classification of test data into normal or attack. KDD Cup'99 dataset was used for training and testing purpose. Due to huge number of instances, a subset of KDD Cup'99 dataset was used for experimentation. It is extremely difficult to process all the features in real time in order to detect network attack and take appropriate actions. On the other hand, not all the features have same relevance for intrusion detection; we need to extract most relevant feature set that can be deployed for efficient detection.

III. PROPOSED WORK

In this Research Work analysis of various outlier detection techniques were used where the algorithms are:

1. K-Mean Algorithm.
2. Genetic and SVM Algorithm.
3. Naïve Baise Algorithm.
4. FPOP Algorithm.
5. SCF Algorithm.

K-Mean Algorithm

K-means clustering finds the centroids, where the coordinate of each centroid is the means of the coordinates of the objects in the cluster and assigns every object to the nearest centroid. The algorithm can be summarized as follows. *k*

Step 1: Select objects randomly. These objects represent initial group centroids. *k*

Step 2: Assign each object to the group that has the closest centroid.

Step 3: When all objects have been assigned, recalculate the positions of the centroids. *k*

Step 4: Repeat Steps 2 and 3 until the centroids no longer move.

Genetic Algorithm for Outlier detection

A Genetic Algorithm (GA) is a programming technique that mimics biological evolution as a problem-solving strategy. It is based on Darwinian's principle of evolution and survival of fittest to optimize a population of candidate solutions towards a predefined fitness .GA uses an evolution and natural selection that uses a chromosome-like data structure and evolve the chromosomes using selection, recombination and

mutation operators. When we use GA for solving various problems three factors will have vital impact on the effectiveness of the algorithm and also of the applications. They are:

i) the fitness function;
ii) the representation of individuals; and
iii) the GA parameters. The determination of these factors often depends on applications and/or implementation

NAÏVE BAISE ALGORITHM

The Bayes rule provides a way to calculate the probability of a hypothesis based on its prior probability. The best hypothesis is the most probable hypothesis, given the observed data *D* plus any initial Knowledge about the prior probabilities of the various hypotheses *h* (*h* is a hypothesis space containing possible target function).Naive baised algorithm is based on the model called probability based model.

FPOP Algorithm

This is the basic measure for identifying outliers. To describe the reasons why identified outliers are abnormal, the itemsets that are not contained in the transaction (it is said that the itemset is contradictive to the transaction) are good candidates for describing the reasons.

SCF Algorithm

A fast outliers identification method for categorical data sets named SCF (Squares of the Complement of the Frequency). The proposed method aims at finding outliers, observations with small marginal frequencies. For each observation, it calculates frequency score named SCF(*xi*):
SCF uses the sum of squares of the complement of the marginal frequency instead sum of the marginal frequency to emphasize the difference between frequent and infrequent categories. In contrast to other outliers identification methods in categorical data sets, it considers number of categories in the categorical variables.

IV. EXPERIMENTAL SETUP

The KDD Cup 99 dataset has been used for implementation of our proposed algorithms. KDD Cup 99 dataset are based on the 1998 DARPA initiative which provide a benchmark for evaluating IDS. In 1998, DARPA Intrusion Detection Evaluation Program was set up and managed by MIT Lincoln Laboratory at MIT. The objective of evaluation program was to evaluate research in the field of intrusion detection. The evaluation program launched a standard dataset i.e. DARPA98 dataset which includes a wide variety of intrusion simulated in a military network environment. DARPA98 dataset was used as training dataset as well as testing dataset to evaluate the performance of IDS. Lincoln Laboratory set up an environment to acquire nine weeks of TCP dump data. DARPA98 dataset contains seven weeks training data and two weeks testing data [15]. The improved version of DARPA98 dataset is termed as KDD Cup 99 dataset. KDD Cup 99 dataset was used for the Third International Knowledge Discovery and Data Mining Tools Competition held in conjunction with KDD-99 the Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections.

DIFFERENT ATTACKS

KDD Cup 99 dataset contains 22 different types of attack. These various attacks has been described in four major categories: Denial of service (DOS) attack, User to root (U2R) attack, Remote to local (R2L) attack, Probes attacks.

Denial of Service (DoS): A DoS attack is a type of attack in which the hacker makes a computing or memory resources too busy or too full to serve legitimate networking requests and hence denying users access to a machine.

Remote to User Attacks (R2L): A remote to user attack is an attack in which a user sends packets to a machine over the internet, which s/he does not have access to in order to expose the machines vulnerabilities and exploit privileges which a local user would have on the computer.

User to Root Attacks (U2R): These attacks are exploitations in which the hacker starts off on the system with a normal user account and attempts to abuse vulnerabilities in the system in order to gain super user privileges.

Probing: Probing is an attack in which the hacker scans a machine or a networking device in order to determine weaknesses or vulnerabilities that may later be exploited so as to compromise the system.

V. RESULT ANALYSIS

All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running windows 8. The discussed feature selection algorithms were implemented using language Java. Proposed as well as existing algorithms were applied one by one in same dataset. At last, comparative study was prepared for all algorithms.

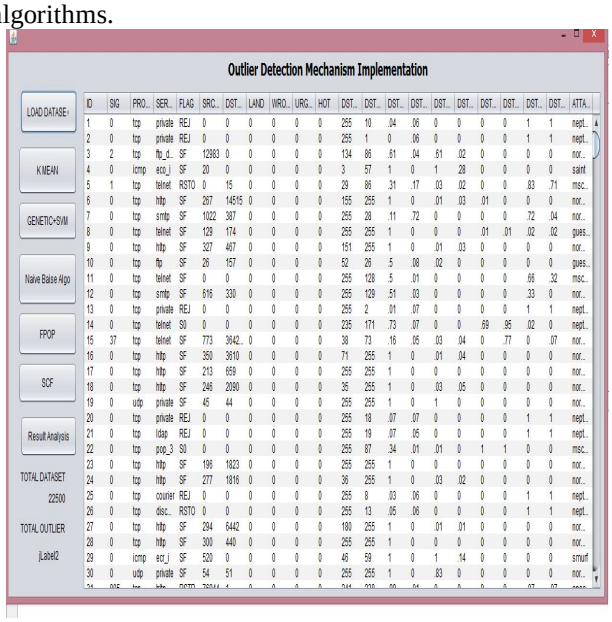

Figure 1: All dataset loaded into our framework.

After loading complete dataset we have performed the outlier detection using different technique on our dataset and got following table data of best 3:

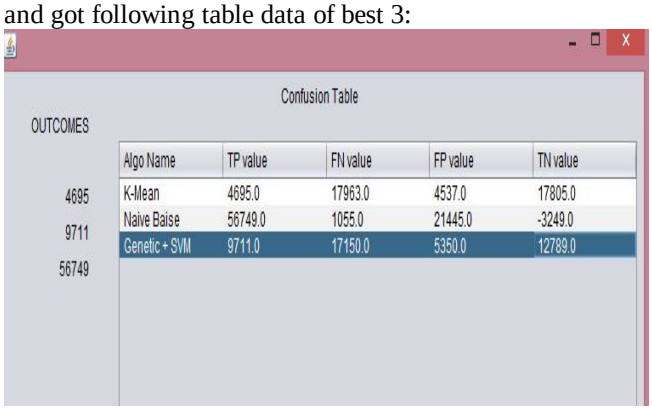

Figure 2 : Confusion table after performing outliers.

Upon based on different algorithm we have calculated four parameters :

1. Accuracy.
2. Detection Rate.
3. Precision.
4. Recall.

And observed following 3 best algorithm results.

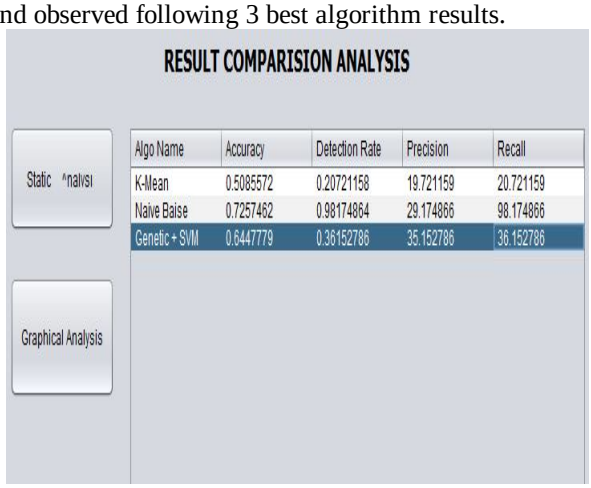

Figure 3: Result Analysis of Best 3 Algorithms.

Upon analyzing various result here thus we can assure about the algorithm which is applied Genetic+SVM perform best among other in the term of accuracy, precision and recall which is better than other two competitive algorithm for IDS.

VI. CONCLUSION AND FUTURE WORK

In this Paper we have conducted various experiment of different algorithm and observed results, by considering all features in dataset for outlier detection. We have analyzed the result from multiple algorithms that select relevant features for the proposed frameworks. A subset of KDD Cup'99 dataset was used for evaluating the performance of system. Also based on our observation we can notify that the hybrid approach defined by us is an effective approach in order to retrieve outlier which was perform on the dataset. In future, we can replace decision tree classification model with some other model and compare its classification accuracy with the proposed framework.

REFERENCE

[1]. Ayman Taha#1, Osman M. Hegazy, "A Proposed Outliers Identification Algorithm for Categorical Data Sets".

[2]. Dewan Md. Farid, Mohammad Zahidur Rahman," Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm" journal of computers, vol. 5, no. 1, january 2010.

[3]. Zengyou He1, Xiaofei Xu1, Joshua Zhexue Huang2, Shengchun Deng," FP-Outlier: Frequent Pattern Based Outlier Detection" 2011.

[4]. Agusti Solanas a,1, Enrique Romero b,2, Sergio Gómez," Feature Selection and Outliers Detection with Genetic Algorithms and Neural Networks".

[5]. A. Koufakou1 E.G. Ortiz1 M. Georgiopoulos1 G.C. Anagnostopoulos2 K.M. Reynolds,"A Scalable and Efficient Outlier Detection Strategy for Categorical Data" 19th IEEE International Conference on Tools with Artificial Intelligence.

[6]. Hans-Peter Kriegel, Peer Kröger, Erich Schubert, Arthur zimeck, "Outlier Detection in Arbitrarily Oriented Subspaces", 2012 IEEE 12th International Conference on Data Mining.

[7]. Ozlem GURUNLU ALMA 1♠, Serdar KURT, Aybars UĞUR, "Genetic Algorithm Based Outlier Detection Using Bayesian Information Criterion in Multiple Regression Models Having Multicollinearity Problems", G.U. Journal of Science 22(3): 141-148 (2009)

[8]. Ayman Taha#1, Osman M. Hegazy, "A Proposed Outliers Identification Algorithm for Categorical Data Sets".

[9]. Ayman Taha, Ali S Hadi, "A General Approach for Automating Outliers Identification in Categorical Data" 2013 IEEE.

[10]. A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional," *Statistical Analysis and Data Mining*, vol. 5, no. 5, pp. 363–387, 2012.

[11]. H. Dai, F. Zhu, E.-P. Lim, and H. H. Pang, "Detecting extreme rank anomalous collections," in Proceedings of the SIAM (SDM), pp. 883– 894, 2012.