IJOSTHE
ISSN: 2349-0772 | Volume 5 Issue 6 December 2018 |www.ijosthe.com

Anomaly Detection using Optimized Features using Genetic Algorithm and MultiEnsemble Classifier

Apoorva Deshpande[1], Ramnaresh Sharma[2]

P.G. Student, Department of Computer Science and Engineering, MPCT, Gwalior, India[1]

Associate Professor, Department of Computer Science and Engineering, MPCT, Gwalior, India[2]

ABSTRACT:
Anomaly detection system plays an important role in network security. Anomaly detection or intrusion detection model is a predictive model used to predict the network data traffic as normal or intrusion. Machine Learning algorithms are used to build accurate models for clustering, classification and prediction. In this paper classification and predictive models for intrusion detection are built by using machine learning classification algorithms namely Random Forest. These algorithms are tested with KDD-99 data set. In this research work the model for anomaly detection is based on normalized reduced feature and multilevel ensemble classifier. The work is performed in divided into two stages. In the first stage data is normalized using mean normalization. In second stage genetic algorithm is used to reduce number of features and further multilevel ensemble classifier is used for classification of data into different attack groups. From result analysis it is analysed that with reduced feature intrusion can be classified more efficiently.

KEYWORDS: Intrusion Detection, Genetic Algorithm, Ensemble Classifier, Multilevel Classifiers.

I. INTRODUCTION

In today's modern computer era intrusion occurs in network in each and every fraction of time. Intrusion occurs with a motive to steal data or to change some useful information from network log data. Intrusion detection system can rationally distinguish between normal and intrusive records. Most existing systems have vulnerabilities that make them vulnerable and untreatable. In addition, intrusion detection technology, which is still considered immature and not a perfect tool against intrusion, has done important research. For network administrators and security experts, this becomes a priority and difficult task. Thus, it can not be replaced by safer systems. The data mining-based IDS can effectively identify data of interest to the user and also predict the results that can be used in the future.

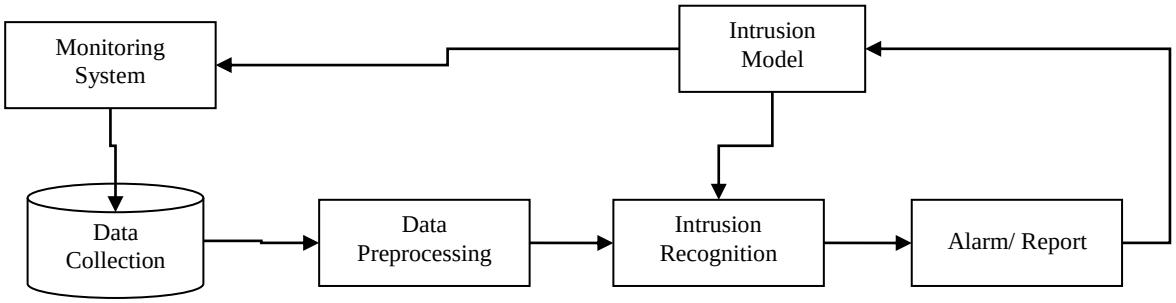

Figure 1: Intrusion Detection System

Fig. 1 illustrates the architecture of IDS. It has been centrally located to capture all incoming packets transmitted on the network. Data is collected and sent to pre-processing to eliminate noise; Irrelevant and missing attributes are overwritten. Thus, the pre-processed data are analysed and classified according to their severity. If the registration is normal, it does not require further changes or sends the report to activate the alerts. Alarms are triggered based on data status so the administrator can handle the situation in advance. The attack is modelled to allow classification of network data. The whole process above will continue as soon as the transfer starts.

Based on the data analysis technique, there are two broad categories of IDS titles, which are mainly based on signatures and anomalies. A signature-based system detects attacks by analyzing network data for attack signatures stored in its database. This type of IDS detects previously known attacks whose signatures are stored in their database. On the other hand, an IDS anomaly appearance - deviations from the traditional behavior of the subjects. The anomaly-based

systems are able to detect new attacks [3-7]. Here are some very common methods used by intruders to take control of computers: Trojan horses, backdoors, denial of service, viruses transmitted via email, package tracking, identity theft and so on. a network package has 42 features and four simulated attacks like [8]:
The classification and selection of features is an important perspective in intrusion detection systems for better performance. Entity classification and feature selection methods are useful to answer the question about the importance of entities in a data set and to classify them into larger or smaller entities. These features help classify network traffic in normal or abnormal (attack) classes. However, features that contribute marginally or unusual to the detection of various types of attacks must be removed to improve the accuracy and speed of intrusion detection systems. The removal of these features will improve the performance of the IDS in terms of calculation, reduction of dimensionality and temporal complexity.

II. RELATED WORK

So far, it has been discussed in this paper about some of the existing approaches which are incorporating IDS. However, there is no universal efficient solution found yet. Each has some limitations. Some of the important contributions in the field of IDS are discussed below in table I.

Table I: Important Contributions in the Field of IDS

Author Name	Approach Used	Conclusion
Sufyan T. Faraj Al-Janabi et al. [1]	The model used BPANN for classification of anomalous network traffic from normal traffic.	93 % (on testing) Accuracy of the system is quite low.
Yinhui Li et al. [2]	K-means clustering is used to compact the dataset into 5 clusters. Ant Colony Optimization (ACO) algorithm was then used to select a small representative subset of the whole dataset. Further Gradually Feature Removal (GFR) is used to reduce the size of the feature set. At the final step, SVM classified the attack instances from benign data.	98.62 %
Feng et al. [3]	Introduced a new classification technique and utilized the advantages of SVM and Clustering based on Self-Organized Ant Colony Network.	94.86 %
Meng et al. [4]	Compared ANN, SVM and DT schemes for anomaly detection in an uniform environment and concluded that J48 algorithm of DT gives better performance than the other two schemes. The detection rate of low frequent attack types (U2R, R2L) was also high.	About 99 %
Manjula C. Belavagi et al. [5]	Classification and predictive models for intrusion detection are built by using machine learning classification algorithms namely Logistic Regression, Gaussian Naive Bayes, Support Vector Machine and Random Forest. Experimental results shows that Random Forest Classifier out performs the other methods in identifying whether the data traffic is normal or an attack.	About 99 %
Saad Mohamed et al. [6]	Presented a hybrid approach to anomaly detection using of K-means clustering and Sequential Minimal Optimization (SMO) classification.	97.36%
Horng et al. [7]	Proposed an IDS based on a combination of BIRCH hierarchical clustering and SVM technique	95.72%
Kuang et al. [8]	Proposed an IDS based on a combination of the SVM model with kernel principal component analysis (KPCA) and genetic algorithm (GA). KPCA was used to reduce the dimensions of feature vectors, whereas GA was employed to optimize the SVM parameters.	95.26%

Khadija Hanifi et al. [9]	In this work, to detect network attacks, used the k-means algorithm a new semi-supervised anomaly detection system.	80.119%.
Wathiq Laftah Al-Yaseen et al. [10]	Presented hybrid SVM and Extreme machine learning technique.	95.75%

III. PROPOSED METHODOLOGY

This section describes the proposed hybrid intrusion detection model. The KDD-99 data set serves as a reference point for evaluating the performance of the proposed model. The flow of the algorithm of the proposed method is described as follows:

Following steps will be used to build the proposed model for intrusion detection:

Step 1: Convert the symbolic attributes protocol, service, and flag to numerical.

Step 2: Normalize data to [0,1].

Step 3: Separate the instances of dataset into two categories: Normal, DOS, R2L, U2R and Probe.

Step 4: Feature Reduction and Extraction.

Step 5: Data Clustering using Kmean clustering

Step 6: The data set is divided as training data and testing data.

Step 7: Train classifier with these new training datasets.

Step 8: Test model with dataset.

Step 9: Finally computing and comparing performance parameters for different classifiers.

Followings Steps are performed in proposed methodology:

A. Data Selection

The first step involves selection of dataset KDD-99 which consists of five classes:

- Normal class
- Four are attack classes known as DoS, U2R, R2L and Probe.

Denial of Service (DOS) is the type of attack that denies legitimate users or waits for resources to be exploited by malicious users so that legitimate users cannot use resources or their resource request is denied. Example: Smurf, Neptune, teardrop, back etc.

In Probing, attackers collect all information on computer networks and look for vulnerabilities to launch the attack. Port scanning is one of the main attacks in this category, the others are ip-sweep, saint and nmap etc.

In, Remote to local (R2L) attackers attack computer systems so that vulnerabilities are accessible as local users. The attacker attempts to create an account on the victim machine by guessing a password or by attacking. Guess password, multi-hop, phf, spy, Warezclient etc. are examples of R2L attacks.

User to root (U2R) with local access to the operating system of the vulnerability system to obtain the root privileges of a system. Example: buffer overflow, root-kit, land module, Perl etc.

B. Data Preprocessing

Data pre-processing was performed manually by deleting the duplicate instances of the KDD-99 dataset and subdividing the instances into different classes. The method starts by removing some redundant instances from the commonly used classes. The result of the preprocessing step provides a compact data set with the elimination of redundancy and imbalance.

C. Data Normalization

After data preprocessing data normalization is performed. Attribute normalization reduces the computational complexity by normalizing the data values between 0 and 1. For this mean range normalization technique is used. Mean range value is calculated as:

$$\text{Data}_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad \text{(i)}$$

Where , x_i = original data of the feature or attribute

$\min(x_i)$= minimum value of data attribute

$\max(x_i)$= maximum value of data attribute

Normally xi is set to zero if the maximum is equal to the minimum.

D. Feature Selection and Reduction

The aim Feature selection phase is to further select only those features from the database which are relevant for proper classification of the dataset and consequently reduces the feature space dimension so as to reduce complexity by removing irrelevant data. This task is accomplished by using genetic algorithm which are discussed below:

The Genetic algorithm operates on binary search space as the chromosomes are bit strings. To begin with gemetic algorithm following steps are performed:

Initial Population Selection: Initially, the genetic algorithm begins with a primary population including random chromosomes that consist of genes with a sequence of 0s or 1s.

Evaluate Fitness Function: In genetic algorithm binary chromosome are employed i.e. '1' and '0'. The gene having gene value '1' is selected feature whereas '0' gene represents that that feature s not selected for evaluation. Out of all features top 'n' features are selected for next generation.

Selection: In each successive generation, a new population is created by selecting the members of the current generation based on their relevance. Regulators are almost always selected, which leads to a preferred selection of the best solution.

Crossover: The most important step in the production of a new generation is the crossover. To create a new generation, the crossover process selects some individuals as parents in the collection determined by the breed selection process.

E. Clustering

K-means cluster is a kind of unsupervised learning, that is employed once you have unlabelled data (i.e., data without outlined classes or groups). The goal of this algorithm is to search out groups within the data, with the quantity of groups described by the variable K. The algorithm works iteratively to assign each data point to at least one of K groups supported the options that are provided.

Algorithmic steps for k-means clustering

Let $X = \{x_1,x_2,x_3,\ldots\ldots,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots,v_c\}$ be the set of centers.

1. Randomly select 'c' cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. Recalculate the new cluster center using:

$$V_i = (\frac{1}{C_i}) \sum_{j=1}^{c_i} x_i \qquad \text{(ii)}$$

where, 'ci' represents the number of data points in ith cluster.

5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3.

F. Intrusion Detection Phase

For intrusion detection or classification dataset MultiEnsemble classifier is used. For MultiEnsemble classifier at all level classifier random forest algorithm is applied i.e. DOS, Probe, U2R, R2L and Normal are classified using RF algorithm (as shown in figure 2).

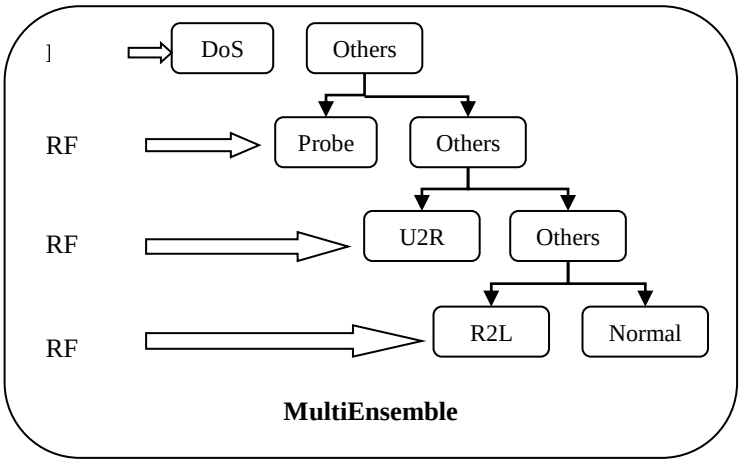

Figure 2: MultiEnsemble Classifier

IV. RESULTS ANALYSIS

For performance evaluation, MultiEnsemble hybrid classifiers are used. The performance evaluation are performed using normalized feature based multilevel classifiers. In this work performance of genetic algorithm based feature reduction technique are evaluated with varying number of features. The result analysis is performed on 10 features, 15 features and 20 features. Table I shows the result analysis.

Table I: Performance Analysis of Detection Rate over Feature Reduction Techniques

Performance Parameters	10 Features	15 Features	20 Features
Accuracy	98.0931	99.9357	99.9893
Detection Rate	98.0912	99.5817	99.9478
FNR	1.9088	0.4183	0.0522
FPR	1.9067	0.0099	0.0043
FAR	1.9077	0.2141	0.0283

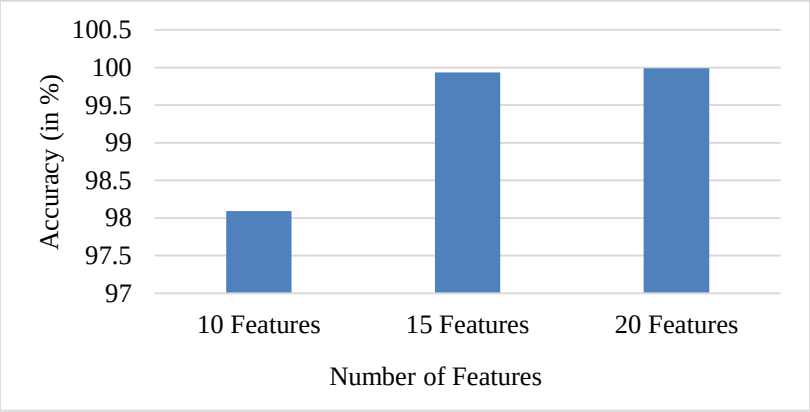

Figure 3: Accuracy Comparison for varying Number of Features

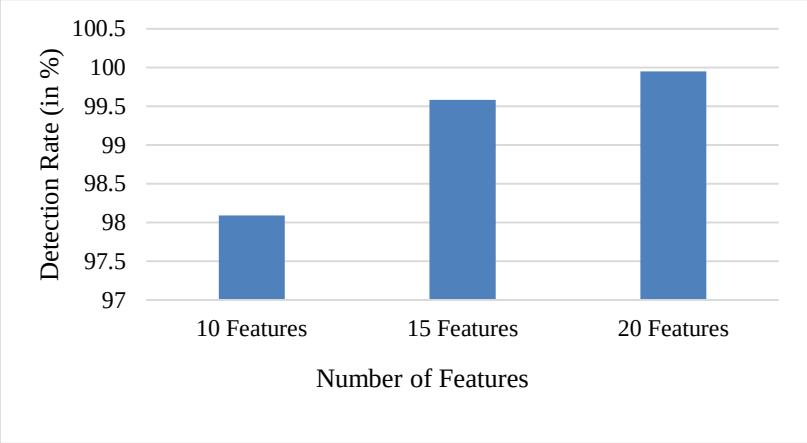

Figure 3: Detection Rate Comparison for varying Number of Features

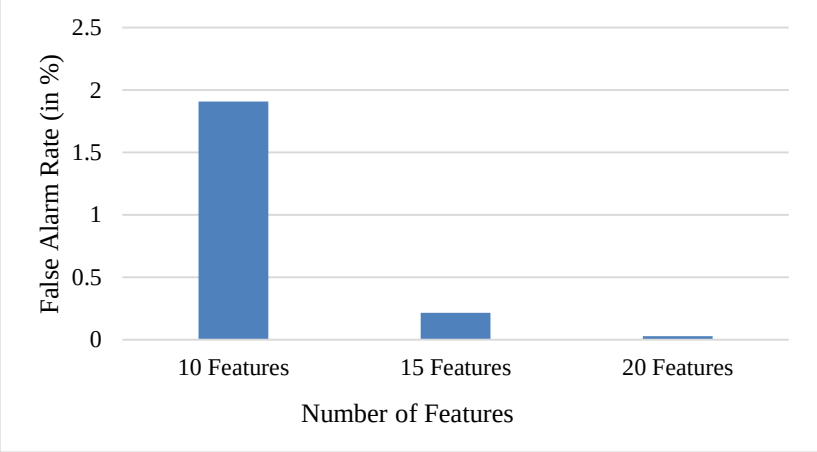

Figure 3: False Alarm Rate Comparison for varying Number of Features

V. CONCLUSION

This paper proposed intrusion detection system that is based on reduced number of features. The system extracts features using concepts of genetic algorithm. The method uses elimination of redundant and irrelevant data from the dataset as well as normalization in order to improve resource utilization and reduce time complexity. A classification system was designed using MultiEnsemble hybrid classification which was trained on KDD99 dataset. From the result analysis it has been analyzed that accuracy, detection rate and false alarm rate of MultiEnsemble classifier outperforms better with 20 features.

REFERENCES

1. Sufyan T Faraj Al-Janabi, Hadeel Amjed Saeed, “A neural network-based anomaly intrusion detection system”, IEEE, 2011.
2. J. Ryan, M. Lin, and R. Miikkulainen, “Intrusion Detection with Neural Networks,” Conference in Neural Information Processing Systems, 943–949.
3. A. K. Ghosh and A. Schwartzbard, “A Study in Using Neural Networks for Anomaly and Misuse Detection,” Conference on USENIX Security Symposium, Volume 8, pp. 12–12, 1999.
4. Meng, Y.-X., “The practice on using machine learning for network anomaly intrusion detection”, International Conference on Machine Learning and Cybernetics (ICMLC), vol. 2, IEEE, 2011.
5. Manjula C. Belavagi and Balachandra Muniyal, “Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection, Procedia Computer Science”, Elsevier, 2016.
6. Saad Mohamed Ali Mohamed Gadal and Rania A. Mokhtar, “Anomaly Detection Approach using Hybrid Algorithm of Data Mining Technique”, International Conference on Communication, Control, Computing and Electronics Engineering, IEEE, 2017.

7. Shi-JinnHorng, Ming-Yang Su, Yuan-Hsin Chen, Tzong-Wann Kao, Rong-Jian Chen, Jui-Lin Lai, Citra Dwi Perkasa, “A novel intrusion detection system based on hierarchical clustering and support vector machines” Expert Systems with Applications, Elsevier, vol. 38, pp. 306–313, 2011.
8. Kuang, F., Xu, W., & Zhang, S., “A novel hybrid KPCA and SVM with GA model for intrusion detection”, Applied Soft Computing Journal, vol. 18, pp. 178–184, 2014.
9. Khadija Hanifi ve Hasan Bank “Network Intrusion Detection Using Machine Learning Anomaly Detection Algorithms” , IEEE, 2016.
10. Wathiq Laftah Al-Yaseen, Zulaiha Ali Othman, Mohd Zakree Ahmad Nazri, “Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System”, International Journal in Expert Systems With Applications, Elsevier, 2017.
11. Sumaiya Thaseen Ikram, Aswani Kumar Cherukuri, “Intrusion detection model using fusion of chi-square feature selection and multi class SVM”, Journal of King Saud University –Computer and Information Sciences, 2016.
12. Yadigar Imamverdiyev “Anomaly detection in network traffic using extreme learning machine”, IEEE, 2016.