IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Classification of Attack for IDS Using Binary Genetic Algorithm Based Feature Selection

S. Rani

Department of Computer Science and Application, Kurukshetra University, Kurukshetra, Haryana, India

*Corresponding Author: savita446@gmail.com

Available online at: www.ijcseonline.org

Accepted: 24/May/2018, Published: 31/May/2018

***Abstract*-** IDS is used to detect any kinds of attacks that may harm the safety of systems. A capable IDS system needs low FAR, and high accuracy. In this paper, we have used fully distinct DM approaches on IDS with the KDD data set. Here the BGA which offers a new method used for fixing normal & DOS.

***Keywords*—**Intrusion Detection System, Binary Genetic Algorithm(BGA), Classifiers, Anomaly Detection

I. INTRODUCTION

21st Century is the era of Technology. Today, the internet is essential for us in every field of education, businesses, banking etc. There is a big organization in the world, the security of the data of these companies is very important. Hardware, software, and network engineers all make these systems together. Intrusion detection system will be used to prevent large-scale attacks, to prevent small attacks and even malicious activities. Dos (Denial of service), R2L (Root To local), Probe etc. are malicious activities which are monitored by IDS.

It has been circulated in many parts: Anomaly detection system, signature-based misuse, Host, Network-based, Stack-based. Generally, it is used in two ways, Host Based Ids (HIDE) and Network-Based Ids (NIDE). Host Based Ids examines the single system, while the use of the Network based ids is for the incoming packets section. IDS are separated in two ways according to the detection: signature-based IDS and anomaly-based IDS. Signature-based Ids contain a database of valid attack and activity is compared with the signature database. If there is any suspicious activity then there is an alarm sound. A loss of signature IDS does not detect the new attack, because new match not found. The anomaly is used to detect informal intrusion, it finds patterns that are not compatible with behavior. There are many applications such as Bio-Information, Fraud Detection, Cyber Security, and Image Processing in which intrusion detection systems are used. DataMining is extracting the valuable knowledge from data. It uses a lot of techniques that detect intrusion such as classification, clustering, frequent pattern, data stream etc. Machine Learning is a subpart of data mining which can describe algorithm designing that can learn and predict the date. Machine learning is of two types: supervisor and unsupervised learning. Clustering & dimensional reduction (PCA, K-Means) is part of unsupervised learning and Regression and Classification (Knn, Svm) is part of supervised learning.

For the time being, many attacks are large, for which the system is being used, the number of an algorithm like numbering algorithm like the Data mining Techniques, which has some challenges in the earlier work, which is like this.

i to select the best classifier to give high accuracy and less false positive rate.
ii. To select best features of records
iii. To overcome false positive rate and high computational complexity
The main goal of intrusion detection is to detect the attack of the future, in which new learning techniques are used. A new system in the proposed work, which is a discovery detection system for detecting new attacks and new networks. This system not only reduces the high detection rate but also the false alarm rate. For this, feature selection technology has been used with the binary genetic algorithm and KDD dataset

Our goal is not only high detection rate for malicious activity in IDS but also to reduce false warning, so this paper has used data mining technology which will reduce false alarm rate [1] In section ii, it has been explained about the tasks behind the back which discuss the problem statement

and the objectives. Section iii describes the proposed work that describes the method, such as How Binary Genetic Algorithm Works on optimal feature Selection Session iv Describes the Result and Discussion about the outcome, which classifier is performing the Better and the section tells about v discuss conclusions and futures work.

II. RELATED WORK

A considerable measure of work has been done in the field of intrusion detection system utilizing different Data mining methods. In this area, various investigations have been checked on and execution of various data mining methods utilizing a KDDCUP99 dataset

R. Samrin [2] Detecting attacks in computer systems and networks is a significant research area. There is a lot of information available for detecting the network intruder. But this is not all that capable of identifying new types of attacks. In this paper, kddcup99 distributed the intrusion on the dataset and New and effective techniques have been implemented for identification.

R.Jakhale[3] In this paper the unknown packets have been detected, for which clustering and sliding window model technique with data mining algorithms have been used. This model has been used on various datasets from different network flows. The result is reasonable, therefore this system provides high recognition capability and preserves very low false alarm rates.it provides a more accurate, effective view.

Sultana et al [4] The AODE Naive Bayes algorithm convection has been used to prevent infiltration attacks in IDS. The proposed model uses AODE algorithm to identify different types of attacks. It detects intelligent network attacks. The proposed model is for lower FAR and higher detection.

H. Om et al [5] this paper has been proposed to detect a hybrid attack system. It combines the k-Means and some Classifiers. In order to detect an attack, entropy-based feature selection has been used, which selects the imported attribute and removes the irredundant data. It is used on kdd data set. This dataset is used around the world for various types of data attack demonstration. This system can detect attack which is divided into four areas, its main goal is to reduce the false alarm rate of the target

.**L. Li et al [6]** Network intrusion detection system is used primarily for security technology and against a strong attack. Data mining or machine learning technology is used in network detection and prevention systems. In view of this, with the continued obstruction of itemsets in the classical uplift algorithm, a length-decreasing was proposed to detect an attack on the basis of data mining, which is a better apriori algorithm**. Goeschel [7]** offers an efficient model. To detect the attack, in order to make the system more efficient, it reduces the false positive rate. This paper proposes to reduce false alarm by using data mining technique such as Svm, DT, and Nb.

N. Shahabad et al [8] has proposed new facility selection process to remove irrelevant features. We have used the proposed and some other survival techniques along with the experimental analysis of the Kdd dataset, DT, and some other classification algorithm.

I. S. Thaseen et al [9] The IDS plays a major role in detecting attacks in computers or networks Anomaly detection is used to detect unnecessary attacks. Although many problems in the traditional direction such as high false alarm rates, low detection rates against new network attacks. This paper proposes the method of collecting SVM by optimizing the kernel parameter using the automatic parameter selection technique. The proposed system shows that this method is successful in identifying attacks.

III. METHODOLOGY

It describes data mining techniques that are applicable in the proposed model. The proposed system convey the Binary Genetic Algorithm make optimized feature selection with Dos and normal in the records training dataset. The resulting feature used for divided of dos and normal activities of testing data. Finally, we show the classification of attack types using classification and comparing the Rf, Svm, and Knn for better results for high detection rate and low false alarm rate

A DATASET READING AND PREPROCESSING

In this, Kdd dataset has been used in which there is an archive of 4,900,000 and these records have been divided into two parts -Train and test dataset. The work on 10% of this dataset has been displayed. Under this data set, 41 Feature is included and there is an output function which tells the normal and abnormal activity. The first step is to read the dataset and read the 2, 3,4,42 column and the next step is to pre-process the variable data change in numerical data.

B FEATURE EXTRACTION & NORMALIZATION

Next step is Feature Extraction which has been performed for reduced the size of the dataset. kdd cups feature is included with different values when these numbers are processed directly, then this is time consummation and the classifiers will not be accurate. So, the process of normalization will be performed. There are many types of Normalization. The commonly used sequence is Z-score, Min-max scaling, and decimal scaling. here used min-max formula which showed the below.

- **Min-Max Normalization**: Min-max normalization [10] performs a linear transformation on the original data. Min-max normalization maps a value of P to d′ in the range [new_min (p), new_max (p)]. The min-max normalization is calculated by the following formula:

$$d = \frac{[d - \min(p) * [new\_max(p) - new\_min(p)] + new\_min(p)}{[\max(p) - \max(p)]} \quad (1)$$

- Where min(p)=minimum code about point,max(p)=max value about point,new_max(p) in upper equation,so formula is

$$d' = \frac{d - \min(p)}{\text{Max}(p) - \min(p)} \quad (2)$$

C. Optimal feature selection with a binary genetic algorithm

In this section binary genetic algorithm with information gain used for the optimized problem. Optimal Solution is used to reduce irrelevant feature as the irrelevant feature less accuracy in learning method. For this, the binary genetic algorithm is used which depends on the best fitness value. And feature selection depends on search technique and evolution. Search technique used to finding new subparts or values and evolution used as the scored value for different feature selection. Example of search is, selection, mutation information, and entropy is an example of evolution in feature selection. Information gain is telling about the information feature which belongs to the class. It based on fitness value. Entropy is a measure of disorder in the dataset. The formulas of IG and Entropy is defined as:

Entropy $(E(s')) = \sum_{i=0}^{k} P_{i \log -}(P_i)$

Entropy for fitness values $E(d, a) = \sum_{j=1}^{q} (g_j/n \times E(dj))$

Information gain $(I(d, a)) = E(d) - E(d\text{-}a)$

Figure 3. 1. Shows Flowchart of proposed work and this flow chart describes the working of binary genetic algorithms, which is based on the best fitness value. In this proposal, binary genetic algorithms have been used for dos and normal

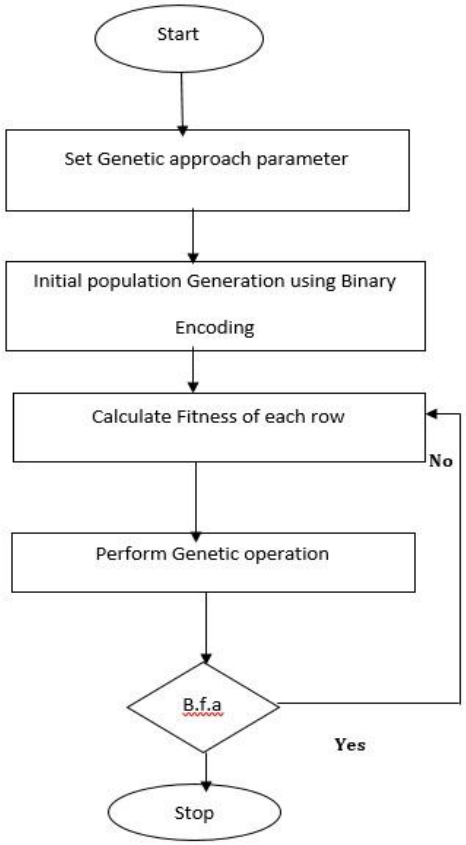

Figure 3 1. Flowchart of proposed work

IV CLASSIFIER

Classification is a technique for data mining that is used to predict target classes, using many mathematical techniques such as decision trees, linear programming, neural networks, static and some algorithms such as rule-based, genetic algorithm, neural network. It is also used as supervisor learning, which classifies unknown and future. Classification Classifier can be divided in 3 ways: one, binary, multiclass value. one class based on signal class classifier.outlier, binary based on two classes eg.in this research paper consists of two value of class Normal and Dos and Multiclass which based on more than two classes eg. Producing revenue. This paper has been compared to 3 classifications, which is an Svm, Knn, and Random forest .two classes For this, 3 steps are described as follows:

Step1: In the first step, the Train, Test Data Set has been taken.

Step2: In the second step, Module will be taken but before it calculates the Accuracy with the predictive value and the confusion matrix.
Step3: In the third step, calculate the detection rate, false rate and time complexity with Higher Accuracy
Here's the use of 3 types of classifications that are described in this way.

- **Knn:** It's a kind of classification algorithm that is also called K Nearest Neighbor.it is a simple algorithm and work as a distance function.Most time usable function.in this paper accuracy of Knn is higher at 98.82%. with high detection rate and low false alarm rate.
- **Random forest:** it consists of many decision trees used as supervisor learning. this paper is second higher accuracy % 95.75 which is better than Svm.
- **Svm**: It is the type of classifier which used as supervises learning.it is last accuracy gain function in this work.

B. PERFORMANCE EVOLUTION

Ids is a monitoring system which used to generate the alarm for malicious attack and Some parameters have been defined for measured performance.

- **True positive (TP)** = the number of cases correctly identified as normal
- **False positive (FP)** = the number of cases incorrectly identified as an attack like Dos
- **True negative (TN)** = the number of cases correctly identified as -ve
- **False negative (FN)** = the number of cases incorrectly identified as –ve
- **Accuracy** is measured as no of correct prediction (TP+TN) is divided by total no of the dataset

V RESULTS AND DISCUSSION

A ENVIRONMENT (DATA, SET\TOOL)

The practical carry on Mat lab R2017b, mixed with many packages dealing with libSVM, C-SVM, and distribution evaluation[11].practical can operate at KDDCUP99[12] dataset. A no of attack can consist of many types of network traffic. Scaling is performed enforce in staring state since the value can change into a numerical pattern transform into scaling form. A detail can be collected as [13]. The dataset contains the 41 attributes where 34 records are continuous and 7 are discrete values. The dataset, contain the 10% which is equal to 20,000. The dataset can categories into two classes normal, dos than it can filter and normalized than optimized feature selection can perform by GA and comparison in SVM, RF, and KNN which is best.

B GRAPH/CHART(KNN/RF/SVM)

Table 4.1 display relationship for Accuracy, Detection Rate, False Alarm Rate, Time Complexity(sec) for ids used techniques for classification where high accuracy, high detection rate and false alarm rate reduced by GA+KNN and time complexity also high for KNN+GA.

Approach	Coparsion b/w Classifier			
	Accuracy %	*Detection Rate*	*False alarm rate*	*Time Complexity (sec)*
GA+KNN	*98.8214*	*98.8214*	1.1786	1.723
GA+RF	94.7571	94.7571	5.2429	26.711
GA+SVM	94.7143	94.7143	5.2857	23.939

TABLE 4. 1 EXECUTION OF GA+(KNN+RF+SVM)

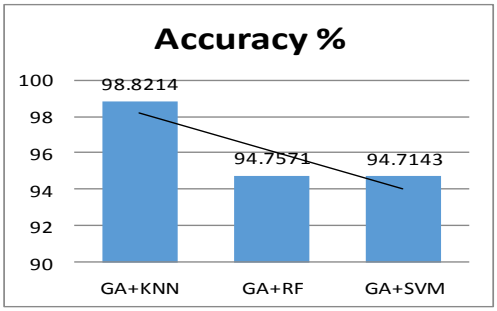

Figure 4.2 shows the accuracy comparisons of the proposed model with a binary Genetic algorithm with feature selection.

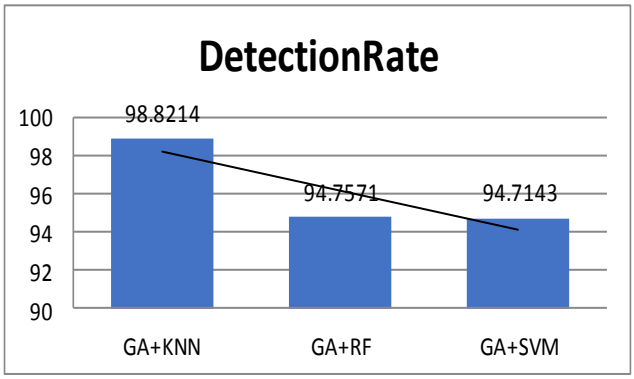

Figure 4.3. Comparison of proposed model by detection ratio

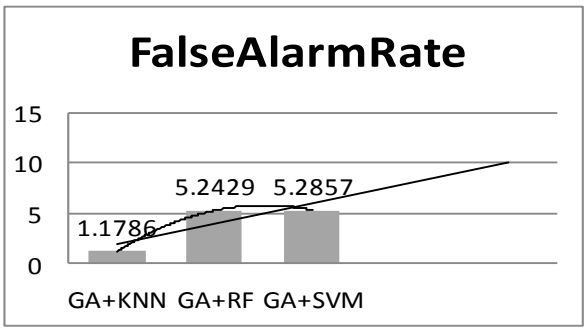

Figure 4. 3 FAR for classifiers

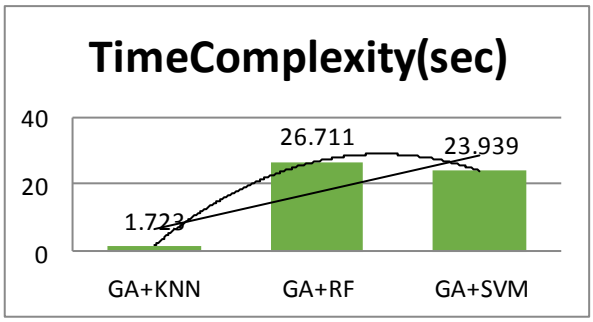

Figure 4. 5. Time complexity in sec

C. PERFORMANCE PARAMETER

- **True positive** TPR=TP/ (TP+FN),
- **False positive rate**: FPR=FP/ (FP+TN) FPR=FP/ (FP+TN),
- **True negative rate**: TNR=TN/ (FP+TN),
- **Confusion matrix**, it means that FPR=FP/ (TP+FP) FPR=FP/ (TP+FP) and FNR=FN/ (TN+FN).
- **FalseAlarmRate** = (fp/ (tn+fp))*100
- **Detection Rate** = (tp/ (tp+fn))*100
- **Accuracy** = ((tp+tn)/ (tp+tn+fp+fn))*100

VI CONCLUSION AND FUTURE WORK

In this paper, we have created new approach as a binary genetic algorithm for ids based on KDD dataset. Filtering, optimal binary genetic algorithm, and classifier have been used for the purpose. The use of filtering and normalized has been used for reducing the dataset and binary genetic for the optimal solution. The classification prediction has been used for accuracy.

Future, one can work on the decision tree, a fusion or hybrid with PSO, ANT for high accuracy, low false alarm rate and high detection rate with more classifiers.

REFERENCES

[1] R. Kaur and M. Bansal, "Multidimensional attacks classification based on genetic algorithm and SVM," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Dehradun, 2016, pp. 561-565.

[2] R. Samrin and D. Vasumathi, "Review on anomaly based network intrusion detection system," *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pp. 141-147

[3] A. R. Jakhale, "Design of anomaly packet detection framework by data mining algorithm for network flow," *2017 International Conference on Computational Intelligence in Data Science (ICC IDS)*, Chennai, 2017, pp. 1-6.

[4] A. Sultana, and M.A. Jabbar, "Intelligent network intrusion detection system using data mining techniques," In the Proceedings of 2ndInternational Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 329-333, 2016.

[5] H. Om and A. Kundu, "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system," *2012 1st International Conference on Recent Advances in Information Technology (RAIT)*, Dhanbad, 2012, pp. 131-136.

[6] L. Li, D. Z. Yang and F. C. Shen," A novel rule-based Intrusion Detection System using data mining", Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on. Vol. 6. IEEE, pp. 169-172, July 2010.

[7] K. Goeschel," Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector 2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016 564 machines, decision trees, and naive Bayes for off-line analysis", SoutheastCon 2016 IEEE, pp. 1-6, Mar 2016

[8] N. Shahadat, I. Hossain, A. Rohman and N. Matin, "Experimental analysis of data mining application for intrusion detection with feature reduction," *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox's Bazar, 2017, pp. 209-216.

[9] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using a fusion of PCA and optimized SVM," *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, Mysore, 2014, pp. 879-884.

[10] W. Lee and S. J. Stolfo. "Datamining approaches for intrusion detection," In Proceedings of the 7th USENIX Security Symposium, San Antonio, TX, January 1998.

[11] http://kdd.ics.uci.edu/databases/kddcup99/task.html

[12] M. Ektefa, S. Memar, F. Sidi and L. S. Affendey, "Intrusion detection using data mining techniques," *2010 International Conference on Information Retrieval & Knowledge Management (CAMP)*, Shah Alam, Selangor, 2010, pp. 200-203

[13] D.Srivastava, L.Bhambu.Data Classification using support vector machine.Journal of Theoretical Applied Information Technology, Vol.12 (1), pp.1-7, 2010.

Authors Profile

Miss.S.Rani pursed B.Tech in Computer Science and Engineering from Kurukshetra University, India. Now pursuing M.Tech from Department of Computer Science and Application, Kurukshetra University, India. She focuses on research area Datamining and Network Security.

.

