International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

# Evaluation Of Efficiency For Intrusion Detection System Using Gini Index C5 Algorithm

**S.Devaraju**

***Abstract*: Security is the critical part in the computers and the networks which connect the computers each other's through network for communication or exchange the data. It is a wide complex to secure the data while transmitting the data between the system/networks. The intrusion detection is a mechanism to protect the data. There are various existing mechanisms for intrusion detection namely neural network, data mining technique, fuzzy logic, statistical technique etc. In this paper, Principal Component Analysis is applied to reduce the features and Gini index C5 algorithm is used to investigate and evaluate the efficiency and false positive rate. The benchmark KDD dataset is used to evaluate the efficiency and minimize the false positive rate using Gini index C5 algorithm and compare with other algorithm which shows significant improvement and to experiment the KDD Dataset to improve the efficiency and minimize the false positive rate using MATLAB software and demonstrated with the KDD dataset.***

***Keywords* : Intrusion Detection, Gini Index C5, KDD Cup, MATLAB**

## I. INTRODUCTION

Network security is a critical problem in computer infrastructures which is very difficult to secure the data from unauthorized users. Protection of computer security is defined against threats to availability, integrity and confidentiality. Two parts of classification in Intrusion Detection Systems (IDS) are: Signature-based and Anomaly-based IDS. Identifies the intrusion by relating with its previous signatures in the database if the signatures are matching with signature log file is known as signature based IDS. The log file comprises known attacks which detect from the networks or computer systems. The anomaly based intrusion or unknown attacks are detected from networks or a computer system which differs from the normal one. IDS are grouped into Host-based or Network-based IDS. Host-based IDS are identified from the network and it is quick to avoid these attacks which occur while connecting external devices. Network-based IDS are detected from the networks. Whenever the systems are communicate with one another the attacks may copy to connected devices [1]. Organized the paper is: Section 2, background study for IDS is conferred. Section 3, discusses proposed techniques, Section 4, define the KDD Dataset Description. Section 5, produce our experimental results & discussion and Section 6 gives conclusion of the proposed research work.

## II. BACKGROUND STUDY

An ID has a conclusive measurement for identifying the intrusion in the network. Various techniques have been anticipated for detecting the intrusions which are statistical methods, neural network, data mining etc.

C-Means Clustering is applied and author used the less dataset. The reduction algorithm is used to minimize the features to increase the detection rate. The drawback is the usage of less dataset for training [2]. Rough Set Neural Network is used to minimize the resources required for attack detection. KDD dataset is applied to tests the data and provide the reasonable results [3].

Correlation coefficient is applied to experiment mathematically to improve the efficiency. The decision tree and KDD dataset is applied to detect the attacks and improve the efficiency. Different groups of input features proposed and shown the experimental results in recurrent neural network which improved efficiency, particularly for R2L attack. Hierarchical hybrid intelligent model is proposed and results shown with efficiency and computational complexity [4,5].

Data mining technique is applied to detect the attacks. The reduction algorithm is applied to reduce the duplicate attributes and Fuzzy C-Means algorithm is applied to discover initial centers which increase the efficiency [6,13,14]. Rule-based classification is applied to detect the intrusions. The evaluated result shows the improvement in efficiency compared with other techniques [12].

## III. PROPOSED TECHNIQUES

The data mining approaches are well suited for large datasets to find the intrusions like large dataset. Data mining approach helps to abstract knowledge from large dataset. Various data mining techniques include regression, classification, association rule, clustering, ID3, ID5, Gini index etc. [6,13]. The Gini Index C5 Algorithm is used to evaluate the dataset. Gini Index C5 Algorithm is a technique for data mining which measures to Information Gain using gini index. Gini index is applied to each attribute of the dataset to create a decision that splits the dataset into smaller subsets. Gini index inspects the information gain which results from defining an attribute to split the dataset.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

**Dr.S.Devaraju,** Department of Computer Science and Applications, Sri Krishna Arts and Science College, Coimbatore-8, Tamil Nadu, India.

The attribute has maximum information gain to generate the decision. Smaller subset will be generated using Gini index algorithm and to create a leaf node for given decision tree [6]. Info Gain for attribute *A* is used to choose the best splitting attribute. The decision tree is build based on the selected highest InfoGain.

$$\text{InfoGain}(A) = \text{Info}(D) - \text{Info}_A(D) \quad (1)$$

Where,

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (2)$$

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{Info}(D_j) \quad (3)$$

Gini Index is applied to measures the divergence between probability distributions and target attributes values.

$$\text{GiniIndex}(D) = \text{Gini}(D) - \sum_{j=1}^{v} p_j \times \text{Gini}(D_j)$$

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2. \quad (4)$$

Chi-squared statistic is a likelihood ratio to determine the information gain statistical significance.

$$G^2(A, D) = 2 \times \ln(2) \times |D| \times \text{InfoGain}(A). \quad (5)$$

The drawback of information gain is solved between attributes and distinct values using gain ration biases.

$$\text{GainRatio}(A) = \frac{\text{InfoGain}(A)}{\text{SplitInfo}_A(D)}$$

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right). \quad (6)$$

Distance measure is calculated with normalizes the gini index criterion.

$$\text{DM}(A) = \frac{\text{Gini}(D)}{-\sum_{j=1}^{v}\sum_{i=1}^{m} p_{ij} \times \log_2(p_{ij})}. \quad (7)$$

## IV. KDD CUP DATASET DESCRIPTION

The KDD dataset is used to experiment the dataset for IDS and which is benchmarking dataset. Four main classes are grouped with different attacks namely DoS, Probe, R2L, and U2R. Every class contains various attacks with millions of records each with 41 features either normal or an attack [7]. 41 features are grouped either continuous or discrete type [9]. Table 1 indicates various attacks of each class.

**Table 1: Attacks List – Class-wise**

| DoS | R2L | Probe | U2R |
|---|---|---|---|
| land | guess_passwd | nmap | loadmodule |
| back | imap | ipsweep | rootkit |
| pod | ftp_write | satan | buffer_overflow |
| neptune | phf | portsweep | perl |
| teardrop | spy | | |
| smurf | multihop | | |
| | warezmaster | | |
| | warezclient | | |

- ❖ DoS is a Denial of Service which deny the authentic requests like flood,
- ❖ R2L is a Remote-to-Local which unapproved access by the remote system like guessing password,
- ❖ Probing which probing other surveillance like port scanning,
- ❖ U2R is an User-to-Root which unapproved access to root license like buffer overflow.

Principal Component Analysis (PCA) is applied to minimize the unnecessary features. Dimensionality reduction technique is a data analysis and compression tool which is well suited for IDS. It is an influential tool for analysis and difficult to find the high dimensions. While reducing the number of dimensions and found the patterns then which can be compressed without any loss of dimensions [4,12]. Feature selection is applied to raise the efficiency and decline the false positive rate. It improves the execution time and speed of the algorithm. After feature selection there are twenty five features are selected which are duration, Protocol_type, service, src_bytes, wrong_fragment, hot, logged_in, num_compromised, is_guest_login, srv_count, serror_rate, srv_serror_rate, rerror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, dst_host_srv_count, dst_host_same_srv_rate, dst_host_di_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate and dst_host_srv_rerror_rate. Table 2 shows the training and testing dataset.

**Table 2: Training and Testing Dataset**

| Classes | Training Dataset | Testing Dataset |
|---|---|---|
| **DoS** | 5000 | 5000 |
| **Normal** | 5000 | 5000 |
| **Probe** | 1377 | 1000 |
| **R2L** | 3000 | 2992 |
| **U2R** | 35 | 35 |
| ***Total*** | **14412** | **14027** |

From KDD dataset there are 5000 signals from DoS, 5000 signals from Normal, 1377 signals from Probe, 3000 signals from R2L and 35 U2R signals classes designated for training the network. 5000 signals from DoS, 5000 signals from Normal, 1000 signals from Probe, 2992 signals from R2L and 35 U2R signals designated for testing the network. Gini Index C5 algorithm is applied to train and test the selected signals.

## V. RESULTS AND DISCUSSION

The Gini Index C5 algorithm is applied to detect the malicious intrusion by evaluating the signals. Each signal contains 41 features with millions of records and every record contains different types of attacks. Features are reduced from 41 features into 25 features using Principal Component Analysis.

The MATLAB software helps to demonstrate the signals and measure the performance and false positive rate [8]. Soft computing mechanism such as data mining techniques or neural network is applied to address the IDS, which performs more accurate and faster. The reduced features are performing better and improve the performance and minimize the FPR [10, 11]. The confusion matrix is used to measure the efficiency and FPR.

| | Categorized as Normal | Categorized as Attack |
|---|---|---|
| Normal | TP | FP |
| Attack | FN | TN |

- ❖ True Positive (TP) which means the no. of instances categorized as normal truly normal.
- ❖ True Negative (TN) which means the no. of instances categorized as attack truly attacks.
- ❖ False Positive (FP) which means the no. of instances categorized as attack truly normal.
- ❖ False Negative (FN) which means the no. of instances categorized as normal truly attack.

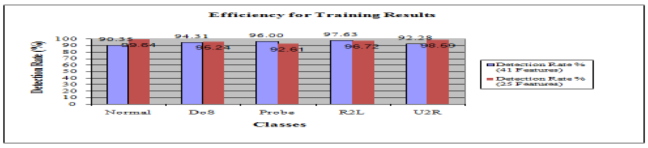

**Figure 2: Efficiency for Testing Results**

The efficiency is measured as follows:

$$Efficiency = \frac{Total\_Detected\_Attack}{Total\_Attacks} * 100 \qquad (8)$$

False Positive Rate (FPR): FPR is ratio among total misclassified instances by the total normal instances.

$$FPR = \frac{Total\ misclassified\ \text{instances}}{Total\ \text{normal instances}} * 100 \qquad (9)$$

### A. Efficiency of Gini Index C5 Algorithm

Proposed system as Gini Index C5-Algorithm has improved the efficiency for DoS, Probe and U2R attacks. In table 3 has shown the efficiency of proposed algorithm.

**Table 3: Efficiency for Gini Index C5 Algorithm**

| Algorithm | No. of Features | Normal | DoS | Probe | R2L | U2R |
|---|---|---|---|---|---|---|
| Gini Index C5 Algorithm (Training Data) | 41 | 90.35 | 94.31 | 96 | **97.6** | 92.28 |
| | 25 | **99.64** | **95.2** | 92.61 | 96.72 | 98.59 |
| Gini Index C5 Algorithm (Testing Data) | 41 | 98.21 | 90.74 | **97.5** | 90.68 | 90.8 |
| | 25 | 98.38 | 90.58 | 93.77 | 96.79 | **99.6** |

Dataset has been experimented with 41 featured and reduced features dataset to measure the efficiency. Depending on the training results, the pictorial depiction of efficiency is shown in figure 1.

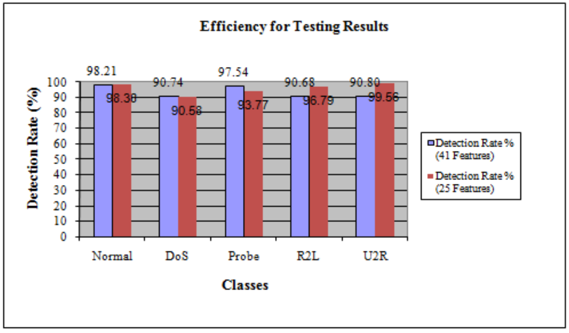

**Figure 1: Efficiency for Training Results**

Depending on the testing results, the pictorial depiction of efficiency is shown in figure 2.

### B. False Positive Rate of Gini Index C5 Algorithm

Proposed Gini Index C5-Algorithm has reduced the FPR. Table 4 shows the false positive.

**Table 4: FPR for Gini Index C5 Algorithm**

| Algorithm | No. of Features | Normal | DoS | Probe | R2L | U2R |
|---|---|---|---|---|---|---|
| Gini Index C5 Algorithm (Training Data) | 41 | 0.040 | 0.024 | 0.057 | 0.077 | 92.28 |
| | 25 | **0.014** | **0.004** | 0.033 | **0.048** | 98.59 |
| Gini Index C5 Algorithm (Testing Data) | 41 | 0.025 | 0.018 | 0.092 | 0.093 | 90.8 |
| | 25 | 0.016 | 0.004 | **0.032** | 0.062 | **99.6** |

Dataset has been experimented with 41 featured and reduced features dataset to measure the FPR. Depending on the training results, the pictorial depiction of FPR is shown in figure 3.

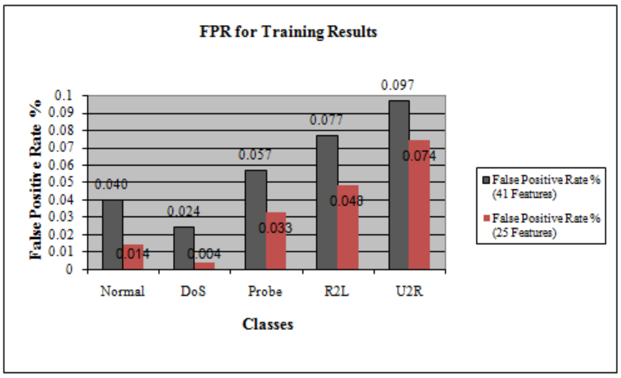

**Figure 3: False Positive Rate for Training Results**

Depending on the testing results, the pictorial depiction of FPR is shown in figure 4.

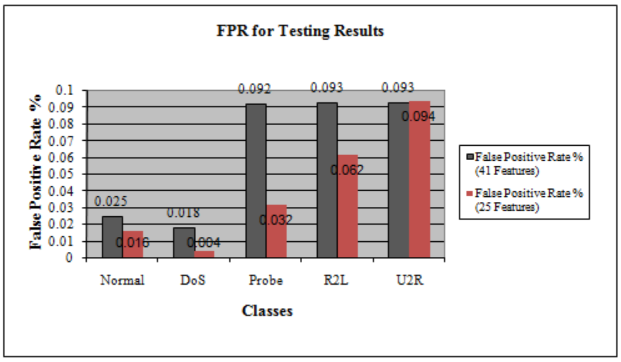

**Figure 4: False Positive Rate for Testing Results**

### C. Comparison of Results

Depends on the experimental results, DR and FPR are compared with existing algorithm. In Table 5 shows the result comparison of efficiency and FPR. The Gini Index C5 Algorithm has significantly improved the efficiency for R2L and U2R and FPR has been significantly reduced for DoS and Probe [1, 11].

**Table 5: Comparison of efficiency with existing Algorithms**

| | | DoS | Probe | R2L | U2R |
|---|---|---|---|---|---|
| Gini Index C5 Algorithm | efficiency | 90.6 | 93.77 | **96.8** | **99.56** |
| | FPR | **0** | **0.032** | 0.06 | 0.094 |
| Layered Conditional Random Fields | efficiency | 97.4 | 98.6 | 29.6 | 86.3 |
| | FPR | 0.07 | 0.91 | 0.35 | 0.05 |
| KDD'99 Winner | efficiency | 97.1 | 83.3 | 8.4 | 13.2 |
| | FPR | 0.3 | 0.6 | 0.01 | 0.003 |
| Multi Classifier | efficiency | 97.3 | 88.7 | 9.6 | 29.8 |
| | FPR | 0.4 | 0.4 | 0.1 | 0.4 |
| Multi Layer Perceptron | efficiency | 97.2 | 88.7 | 5.6 | 13.2 |
| | FPR | 0.3 | 0.4 | 0.01 | 0.05 |
| Gaussian Classifier | efficiency | 82.4 | 90.2 | 9.6 | 22.8 |
| | FPR | 0.9 | 11.3 | 0.1 | 0.5 |
| K-Means Clustering | efficiency | 97.3 | 87.6 | 6.4 | 29.8 |
| | FPR | 0.4 | 2.6 | 0.1 | 0.4 |
| Nearest Cluster Algorithm | efficiency | 97.1 | 88.8 | 3.4 | 2.2 |
| | FPR | 0.3 | 0.5 | 0.01 | 6E-04 |
| Incremental Radial Basis Function | efficiency | 73 | 93.2 | 5.9 | 6.1 |
| | FPR | 0.2 | 18.8 | 0.3 | 0.04 |
| Fuzzy ARTMAP | efficiency | 97 | 77.2 | 3.7 | 6.1 |
| | FPR | 0.3 | 0.2 | 0 | 0.001 |
| C4.5 | efficiency | 97 | 80.8 | 4.6 | 1.8 |
| | FPR | 0.3 | 0.7 | 0.01 | 0.002 |
| Support Vector Machine | efficiency | 91.6 | 36.65 | 22 | 12 |
| | FPR | 0.98 | 2.3 | 11.5 | 7.12 |

## VI. CONCLUSION

An ID is a software application which monitors the kind of activity over the networks. In proposed system, Gini Index C5 Algorithm improves efficiency for R2L and U2R class and reduces the FPR for DoS and Probe. Gini Index C5 algorithm performs significantly to improve the efficiency for R2L is 96.79% and U2R is 99.56% and reduces the FPR for DoS is 0.004% and Probe is 0.032% compared with existing algorithms. MATLAB software is used to demonstrate the dataset using Gini Index C5 algorithm. Hence, it is proposed to consider Gini Index C5 algorithm which significantly improved the efficiency and FPR. The scope for future research includes the extended Gini Index C5 Algorithm for extracting fewer features to apply the other datasets which increase the efficiency and minimize the false positive rate.

## REFERENCES

1. Devaraju S. & Ramakrishnan S., "Performance Comparison of Intrusion Detection System using Various Techniques – A Review", ICTACT J on Comm Tech, Sep 2013, vol.4, iss.3, pp.802-812.
2. Enamul Kabir, Jiankun Hu, Hua Wang & Guangping Zhuo, "A novel statistical technique for intrusion detection systems", Elsevier- Future Generation Computer Systems, 2018, vol.79, pp.303-318.
3. Sandhya Peddabachigari, Ajith Abraham, Crina Grosan, Johnson Thomas, "Modeling intrusion detection system using hybrid intelligent systems", Published by Elsevier Ltd, 2005, pp.1-20.
4. Mansour Sheikhan and Amir Ali Sha'bani, "Fast Neural Intrusion Detection System Based on Hidden Weight Optimization Algorithm and Feature Selection", World Applied Sciences J 7(Special Issue of Comp & IT): 2009, 45-53.
5. Jiankun Hu, Xinghuo Yu, D. Qiu, Hsiao-Hwa Chen, "A simple and efficient hidden Markov model scheme for host- based anomaly intrusion detection", Journal IEEE Netw, January/February 2009, vol. 23 iss. 1.
6. Mei Jiang, Xindan Gan, Chaofeng Wang, Zhuo Wang, "Research of the Intrusion Detection Model Based on Data Mining", Elsevier Energy Procedia, 2011, iss.13, pp.855-863.
7. KDD Cup 1999 Intrusion Detection Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 2010.
8. MATLAB (MATrix Laboratory) tutorials, http://terpconnect.umd.edu/~nsw/ench250/matlab.html

9. Devaraju S., Ramakrishnan S., ”Performance Analysis of Intrusion Detection System Using Various Neural Network Classifiers”, IEEE Proc of the Int Conf on Recent Trends in Info Tech (ICRTIT 2011), MIT, Anna University, Chennai, India. 3-5, June 2011.
10. Devaraju S. & Ramakrishnan S., “Detection of Accuracy for Intrusion Detection System using Neural Network Classifier”, Int J of Emerging Tech and Adv Engg, ISSN 2250-2459 (Online), January 2013, vol.3, Spl Iss.1, pp.338-345.
11. Devaraju S. & Ramakrishnan S., “Performance Comparison for Intrusion Detection System using Neural Network with KDD Dataset”, ICTACT J on Soft Comp, April 2014, vol.4, iss.3, pp.743-752.
12. Ramakrishnan S., & Devaraju S. (2017). “Attack’s Feature Selection-Based Network Intrusion Detection System Using Fuzzy Control Language”, Springer-International Journal of Fuzzy Systems, 2017, vol.19, iss.2, pp.316-328.
13. Minjie Wang, Anqing Zhao, “Investigations of Intrusion Detection Based on Data Mining”, Springer Recent Advances in Comp Sci and Info Engg Lecture Notes in Electrical Engineering, 2012, vol.124, pp.275-279.
14. Devaraju S. & Ramakrishnan S. (2015). “Detection of Attacks for IDS using Association Rule Mining Algorithm”. IETE Journal of Research, vol.61, iss.6, pp. 624-633.

## AUTHORS PROFILE

**Dr. S.Devaraju** received the B.Sc degree in Chemistry in 1997 from the University of Madras, Chennai, and the M.C.A. degree in Computer Applications in 2001 from the Periyar University, Salem, and the M.Phil. degree in Computer Science in 2004 from Periyar University, Salem and also received M.B.A. degree in Human Resource from Madurai Kamaraj University, Madurai in 2007. He received his Ph.D degree in Science and Humanities from Anna University, Chnnai in 2017.

He has 16+ years of teaching experience and 2 years industry experience. He is an Associate Professor, Department of Computer Science and Applications, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu, India. Dr.S.Devaraju is an Reviewer for various reputed Journals and Conferences. He has published more than 10 papers in international journals and conference proceedings. His area of research includes Network Security, Intrusion Detection, Soft Computing, and Wireless Communication.

*Retrieval Number F8593088619/2019©BEIESP*
*DOI: 10.35940/ijeat.F8593.088619*
*Journal Website: www.ijeat.org*

*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication*