A Multi Approach for the Analysis of Feature Selection using Data Gain and BAT Techniques on the Anomaly Detection

Pujala Nanda Kishore[1], Cholla Ravindra Raman[2], Jetti Kumar Raja[3] , Reddy Veeramohana Rao [4]

[1]Department of CSE, Bapatla Engineering College, Bapatla, Guntur, Andhra Pradesh, India.
[2]Department of CSE, Bapatla Engineering College, Bapatla, Guntur, Andhra Pradesh, India.
[3]Department of CSE, Bapatla Engineering College, Bapatla, Guntur, Andhra Pradesh, India.
[4]Department of CSE, Bapatla Engineering College, Bapatla, Guntur, Andhra Pradesh, India.

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Every day, millions of people in many institutions communicate with each other on the Internet. The past two decades have witnessed unprecedented levels of Internet use by people around the world. Almost alongside these rapid developments in the internet space, an ever increasing incidence of attacks carried out on the internet has been consistently reported every minute. In such a difficult environment, Anomaly Detection Systems (ADS) play an important role in monitoring and analyzing daily internet activities for security breaches and threats. However, the analytical data routinely generated from computer networks are usually of enormous size and of little use. This creates a major challenge for ADSs, who must examine all the functionality of a certain dataset to identify intrusive patterns. The selection of features is an important factor in modeling anomaly-based intrusion detection systems. An irrelevant characteristic can lead to overfitting which in turn negatively affects the modeling power of classification algorithms. The objective of this study is to analyze and select the most discriminating input characteristics for the construction of efficient and computationally efficient schemes for an ADS. In the first step, a heuristic algorithm called IG-BA is proposed for dimensionality reduction by selecting the optimal subset based on the concept of entropy. Then, the relevant and meaningful features are selected, before implementing Number of Classifiers which includes: (1) An irrelevant feature can lead to overfitting which in turn negatively affects the modeling power of the classification algorithms. Experiment was done on CICIDS-2017 dataset by applying (1) Random Forest (RF), (2) Bayes Network (BN), (3) Naive Bayes (NB), (4) J48 and (5) Random Tree (RT) with results showing better detection precision and faster execution time. The proposed heuristic algorithm outperforms the existing ones as it is more accurate in detection as well as faster. However, Random Forest algorithm emerges as the best classifier for feature selection technique and scores over others by virtue of its accuracy in optimal selection of features.

Keywords: Functional selection of intrusion detection systems (IDS), information gain, BAT classifier algorithm

1. Introduction

Millions of people in organizations on all continents communicate with each other over the Internet. In the past two decades, the number of people using the Internet has grown exponentially. Currently, nearly 4 billion users worldwide use the Internet [3]. The intrusion detection system (IDS) monitors network traffic to identify malicious events or violations of privacy, and sends alerts to monitoring stations or takes preventive measures against detected threats. IDS can be divided into two categories: one is based on the location of the network installation, or through the detection method shown in Figure 1.

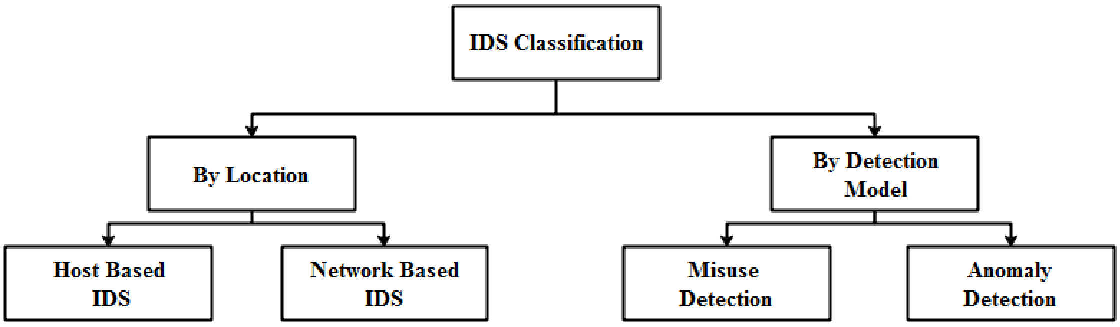

Fig. 1. Classification of Intrusion Detection System (IDS)

Host-based IDS: It runs directly on the client PC and starts to check log files, running processes and client connections. If you need to make changes in important files of the user or operating system, a warning will be sent to the administrator, asking you to take appropriate measures [1].

Network-based IDS: This system monitors and inspects packets of data transmitted over the network to detect actions such as denial of service [1,2]. By the way they are detected, IDS can also be divided into two types: abuse detection and violation detection. Misuse detection can be accomplished by comparing client activity against a stored library of known attack signatures. If it matches, it will check the incoming connection to the saved knowledge base and then stop and block the connection. This type has a high accuracy in detecting known attacks.

Anomaly Detection identifies failures by monitoring abnormal network traffic behavior that could indicate an attack. Abnormal behavior can be defined as a boundary violation, identified as a recurring event on the connection, or it can be defined as a violation of the actual configuration file created by the client for normal behavior. This method can be described as a method based on statistics, data mining and training [4]. Anomaly-based IDS can detect known attacks and new attacks [6].However, the anomaly-based method analyzes the data based on its general attributes (such as size, connection time, and number of packets). Therefore, there is no need to view the content of the message. It can also analyze encryption protocols.

Because of all these advantages, anomaly detection techniques are widely used to detect and prevent network attacks. Anomaly-based IDS can detect known attacks and new attacks [6].

Thus, he does not need to see the content of the message. It can also analyze encrypted protocols. Due to all these advantages, the anomaly detection method is used extensively to detect and prevent network attacks. Previous works [9] - [13] have focused on the application of feature selection techniques in making more accurate identification of anomalies. Previous researchers have always relied upon Information gain for analysis of significant and relevant characteristics. In this study, a version of CICIDS-2017 dataset having critical features has been applied as it demonstrates highly dense traffic and possesses the capabilities to employ huge number of methods at detecting anomalies. As mentioned in [5], the learning model is affected by application of data having multiple features leading to overfit that results in decreased performance, more memory and high computation expenses. But wherever there is involvement of complex functionalities with less values, information gain tend to be supportive. Here, a new mechanism has been introduced to select ensemble features, before slotting them in categories as per their weight values. Then the five classification algorithms, namely, J48 classifier, Naive Bayes classifier (NBC) classifier Bayes Net (BNC) classifier, Random Tree (RTC) classifier and Random Forest (RFC) classifier are assigned filters by each group of entities for detecting anomalies as well as fending off attacks on the dataset. Most relevant and significant features are extracted into different entity groups that are validated after doing comparison of detection results. With more accuracy in detection results, the perception and choice about the important and relevant the feature groups is made. The weighted features which are used in information gain versus anomaly / attack detection method are used to check the relevant and significant features of the selected entity groups. The better precision results shows the features groups which are more relevant and significant. Such features are applied to various classifiers like J48 classifier, Naive Bayes classifier (NBC) classifier Bayes Net (BNC) classifier, Random Tree (RTC) classifier and Random Forest (RFC) classifier on the given data set. Finally the results are validated for relevant and significant features. The ones with better accuracy in detection results tend to be looked up as more meaningful and relevant the feature groups.

In section 2 relevant research contributions made so far on this topic has been presented. In section 3, a brief discussion on the dataset and experimental setup are mentioned clearly.The experimental part, including the results and conclusions of this study has been discussed. Finally, in section 5, the conclusion and potential future work has been discussed.

2. Related works

Recently, most applications depends on the network or computer system and their behavior is to be analyzed and threaten by the known technique called Intrusion detection. Moreover, such technique also interrupt the features of the network or computer system which includes integrity accessibility, and confidentiality of concerned data [5]. The study the characteristics related to the network traffic and also identified number of mechanisms to handle introduction mostly they were filtered, wrapper, and combination of both algorithms [8].However, feature extraction with ensemble of fitter and wrapper assign weight for the every feature and maximum ranked features applied to clustering approach [15]. In some work, most popular resampled method called synthetic minority oversampling technique (SMOTE)[14] is applied to remove class imbalance problem. Later combined two techniques one is the Selection of Ensemble Characteristics (EFS) and the Principal Component Analysis (PCA) and then applied to the AdaBoost-based IDS to improve the performance of classification.One of the most popular wrapper method used by the most of researchers known as information gain (IG) used as a feature selection mechanism and is worked to find the minimum ranking score for each feature as a result set. Next, the ranking weights are used to determine optimal features and are to be considered as final class label. Number of researchers use weight score >0.4, > 0.001 and > 0.8 respectively [16] [14].

3. Feature selection

The mechanism used to extract important and relevant information is known as feature selection. Generally such kind of technique is used to discriminate the class label into relevant and irrelevant functionality .The relevant functionalities had information which is optimal to class and where as in non-informative functionalities the class gained very little information about class [1]. The main objective of feature selection is to filter non-informative features and identify informative features and to pass maximum information related to class output. To achieve this, number of feature selection method are available but generally which is classified into filter, wrapper and combined or ensemble approaches [17][19]. The Filtering method, is one used to access and extract relevant features from the given data using statistical approach. However, in case of the wrapper method selection of the relevant subset of features can be done by using the classification criteria. But the wrapper method is computationally very expensive. The next, method is ensemble or integrated method used to apply feature selection with learning criteria to extract optimal features to the given data. Such kind of ensemble feature selection methods are less expensive compare to the wrapper method.

3.1. Information Gain (IG):

The well-known popular type of filter approach, called Information Gain in which the evaluationof each functionality is depend on how much amount of information is used to identify the desired type of the class attack.

Consider, F is a feature and corresponding class is to be represented as □ and the entropy of the given class related to the feature F is represented as:

$$\mathrm{H}\square = -\sum_{c \in \square} \mathrm{P}(c) \log_2 \mathrm{P}(c) \qquad (1)$$

$$\mathrm{H}\square \mid F = -\sum_{f \in F} \mathrm{P}(f) \sum_{c \in \square} \mathrm{P}(c \mid f) \log_2 \mathrm{P}(c \mid f) \qquad (2)$$

Next, from the (1) and (2) the corresponding Information Gain related to function F to be considered as:

$$\begin{aligned} \mathrm{I}G &= \mathrm{H}(\square) - \mathrm{H}(\square \mid F), \\ \mathrm{I}G &= \mathrm{H}(F) - \mathrm{H}(F \mid \square), \\ \mathrm{I}G &= H(F) + \mathrm{H}(\square) - \mathrm{H}(F \mid \square) \end{aligned} \qquad (3)$$

After calculation of IG all the entities are ordered depend on the calculated $\mathrm{I}G$ value. Finally total M features are to be considered as feature subset with relevant informative feature. Moreover, the resultant features along with $\mathrm{I}G$ value is to be provided suitable information and is helped to find the target output class.

3.2. Bat Algorithm (BA)

The bat algorithm[19-21]is derived from the motivation of the microbats behavior in the field of computational intelligence and optimization .Let consider, every bat flies with random speed to be represented as V_i^t at a desired location to be mentioned as X_i^t having the frequency F_i at iteration t and the solution space represented as d^* .

From the n bats in the population, solution X_* to be calculated with the iterative process. Next, [], the location X_i^t and speed V_i^t are to be updated at the time step t and is to be calculated as:

$$F_i = F_{MIN} + (F_{MAX} - F_{MIN})\beta \qquad (4)$$

$$V_i^t = V_i^{t-1} + (X_i^{t-1} - X_*)F \qquad (5)$$

$$X_i^t = X_i^{t-1} + V_i^t \qquad (6)$$

From the (5) β [0,1] to be a random vector and is to be derived from the uniform distribution.

By applying the local search the solution is derived and then a new solution related to each bat is calculated using the random walk and is to be represented as:

$$X_{NEW} = X_{OLD} + \in A^{t} \qquad (7)$$

Where $\in$ is an error and is random vector derived from the uniform distribution or Gaussian distribution of the range [-1, 1]. Next, A^{t} to be considered as mean value of the all bats at time scale of t. Similarly the loudness A_i^{t+1} and the pulse emission rate r_i^{t+1} are updated as follows:

$$A_i^{t+1} = \alpha A_i^{t} \qquad (8)$$

$$r_i^{t+1} = r_i^{0}(1 - e^{-\gamma t}) \qquad (9)$$

From (8) & (9) $0 < \alpha < 1$ and $\gamma > 0$ are the constants.

4. Proposed method

Machine Learning (ML) based methods are become popular now and are used in this study to improve performance of the Anomaly Detection System (ADS) and also worked for solution to prevent attack from the providers. Ensemble optimization ML based feature selection method applied first and extracted optimal features and then set of classifiers used to detect the attack type. The approach is used a10-fold cross-validation (CV) during the experiment and to validate the model performance. Finally model is to classify attack especially benign traffic attack. The proposed method framework shown in Figure 2, and overall work is divided into major four parts and are given below:

1. **Preprocessing:** The step in which original or raw data is to be converted into desired formats which are helps for further analysis.
2. **Feature Selection :**The second step, applied proposed the IG-BA based feature selection approach used to retrieve the subset of date sets and retrieved most relevant or suitable features related to each type of the attack class.
3. **Classification: The last step of the proposed work is deal classification which is helps to improve overall performance of the** IDS. The number of classifiers used in this work which includes : (i) Random Forest(RF) (ii) Random Tree (iii) naïve Bayes (iv) Bayesian Network and (v) J48.

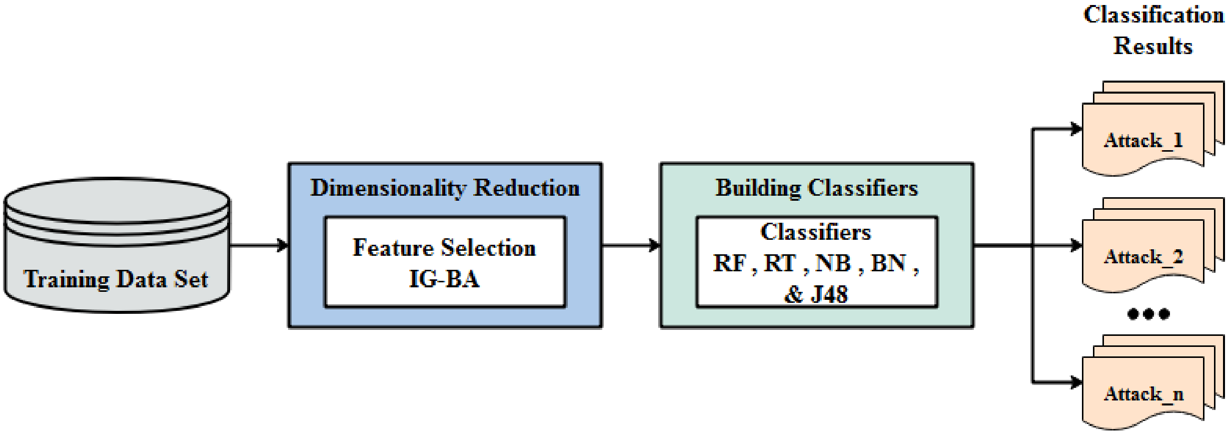

Fig 2: Proposed method framework for the classification

4.1. IG-BA approach for feature selection

The proposed ensemble feature selection approach is called IG-BA method used to evaluate and identify subset of features based on the weighted rank result important features. The IG-BA method is worked first on feature selection based on weighting criteria and derived subset represented as S with distinct k characteristics using the method IG.The method is very simple and derived subset of best features according weighting criteria. However, selected features all the time may not be considered as better features as per the redundancy among the features.

The problem of redundancy among features and also to work on the dimensionality reduction proposed method introduced BA algorithm as an additional step to the feature selection. The feature selection using IG-BA approach is presented in Algorithm1. In the proposed method, first step is population initialization. Later, applied set of rules for updating and helps to move the bats in the population to the research space. In order to find the best solution the BA uses the search concept based on the local random walk. Next, relevant feature subset is derived using IG and produced new solution after updation of both loudness A_i^{t+1} and the pulse emission rate r_i^{t+1}.The process is repeated until get X_{Best} till the end of iterations.

Algorithm: Feature Selection using Proposed IG-BA Method

Input: X_D^i The original data set

Output: X_{Best} The final feature sub set

```
Consider the elements of data set into number of population of $n$ bats
$X_1, ........, X_i (i = 1, 2, ...., n)$ with the speed $V_i$
Let the frequency denoted as $F_i$ , rate of emission to be $r_i$ and finally volume $A_i^t$
Temporary measures : $Fit(X_i)$, $Fit_{temp}(i)$, $X_{temp}(i)$ and $X_{Best}$
while $1 \le t \le Max$
    for $i = 1$ to $n$
        Calculate $F_i$ using (4)
        Next, Update the values of both $X_i$ and $V_i$ using (5) and (6)
        if $r_i^t < rand(0,1)$ then
            Find $X_{Best}$ using IG( $X_i$ )
            Derived new $X_{New}$ using (7)
        endif
        Estimate $Fit(X_{New})$
        if $Fit(X_i) \le Fit(X_{New})$ and $N(0,1) < A_i^t$ then
            $Fit_{temp}(i) \leftarrow Fit(X_{New})$
            Decrease $A_i^t$ and increase $r_i^t$
        endif
        if $Fit(X_{New}) \ge Max$ of $Fit_{temp}$ then
            $X_{Best} \leftarrow X_{New}$
        endif
    endfor
endwhile
$t = t + 1$
end
```

4.2. Classification algorithm

Although several previous works have supported many diverse algorithms, in this work, number of classifiers used which includes: (i) Random Forest (RF) (ii) Random Tree (iii) naïve Bayes (iv) Bayesian Network and (v) J48.

4.2.1. Naive Bayes (NB)

The classification algorithms used to predict probability of a class using Bayes' theorem in terms of statisticalclassification. In some exist works [26-27]it's clear that the impact of one attribute values related to the given class is not influenced on value of other attribute.

4.2.2. Bayes Network (BN)

The model in which among variables there exist encoding probabilistic relation which is called the Bayesian Network (BN). On the general assumption of the behavior of the target system model, the precision of the method is determined, with any notable departure from it is likely to reduce precision in detection. Bayesian networks have been applied in a few anomaly detection studies[22][25].

4.2.3. Random Forest (RF)

Random forest, one of the classification methods, is a classifier in a collection of numbers in a decision tree. Next comes the word Forest, represented as a set of classifiers. Decision trees differ from each other depending on the random selection of the desired attributes corresponding to each node. A number of works related to the detection of anomalies using a random forest have been performed [22] [24].

4.2.4. Random tree (RT)

A decision tree, which is a collection of random attributes called a random tree and a complete tree, is built from a combination of two nodes and branches. However, a node should be viewed as a test attribute and a branch. Decision tables display the final decision after calculating all attributes as class labels. This method has been included in some anomaly detection studies [28] [30].

4.2.5. J48

A machine learning algorithm corresponds to family of decision tree i.e., J48 or C4.5, make use of training data to a decision tree using entropy [43]. Unlike IDE3, this method used to create a decision tree keeping the abilitytogeneratesequence of attributes. The J48 algorithm applied to anomaly detection included in many research work[29].

5. Experimental setup

5.1. CICIDS2017 dataset

The dataset [5], is introduced in 2018 at the Canadian Institute for Cybersecurity and is used to detect DDoS attacks. However, data set is present benign and attack processconsidering real world network traffic data. Also, data set includes 79 features which is comprise of class labels and are used to specify major attacks mentioned: (i) Brute Force SSH (ii) Brute Force FTP (iii) Infiltration (iv) Heartbleed (v) Web Attack (vi) DoS (vii) Botnet and (viii) DDoS and the complete attacks information shown in Table 3. Total 225,746 records related to DDoS and Benign attacks included in CICIDS2017 and each record comprised with total 80 features like (i) protocol (ii) stream ID (iii) source IP (iv) destination IP (v) source port, and etc. The complete records and features is included in Table 1.

Table 1: The records in data set CICIDS2017

Source IP	Source port	Destination port		Duration of flow	Total number of Fwd packages	Total back packets
192.xxx.xx.20	41938	334	...	143346	46	70
192.xxx.xx.20	42978	80		40907	1	1
192.xxx.xx.20	41955	445		143896	47	69
192.xxx.xx.21	12887	54		314	2	2
192.xxx.xx.20	41946	444		142609	44	59
192.xxx.xx.21	33065	55		255	2	2
192.xxx.xx.20	41942	443		142488	47	57
192.xxx.xx.20	41939	444		23838	28	32

5.2. Experimental setup

As an initial model fitting, the complete original data is split into two subsets one is training data (80%) and other is test data (20%). Next, applied proposed IG-BA feature selection method and extracted optimal set of feature set. The algorithm which helps to avoid irrelevant features from the data set and also improved the performance of classification.

Table 2: Training and testing of the CICIDS2017 dataset

Attack class	**No. Records**	**Train set (80%)**	**Test set (20%)**
Benign	61562	49250	12312
Bot	1966	1573	393
Brute force	1507	1206	301
DoS / DDoS	58134	46507	11627
Golden Eye back	10293	8234	2059
Back Hulk	10486	8389	2097
Slowhttptest back	5499	4399	1100
Slowloris back	5796	4637	1159
FTP-Patator	7938	6350	1588
Heartbleed	11	9	2
Infiltration	36	29	7
PortScan	60294	48235	12059
SQL	21	17	4
SSH-Patator	5897	4718	1179
XSS	652	522	130
Total	**230092**	**184074**	**46018**

After performing the feature selection using hybrid proposed method the result subset is applied to different classifiers which are (i) Random Forest(RF) (ii) Random Tree (iii) naïve Bayes (iv) Bayesian Network and (v) J48.

Table 3: Attacks worked on this job

Attack number	**Attack name**
Attack-1	DoS / DDoS attack
Attack-2	Port scan attack
Attack-3	Bot attack
Attack-4	Web attack
Attack-5	Infiltration
Attack-6	Brute force

5.3. Experimental results

The proposed feature selection IG-BA hybrid method applied initially and result subset with important features and then classification algorithms applied over the data with the benchmarks which includes (i) The True Positive Rate (TPR),(ii) The False Rates Positive (FPR), (iii) Precision, and (iv) Recall are used.

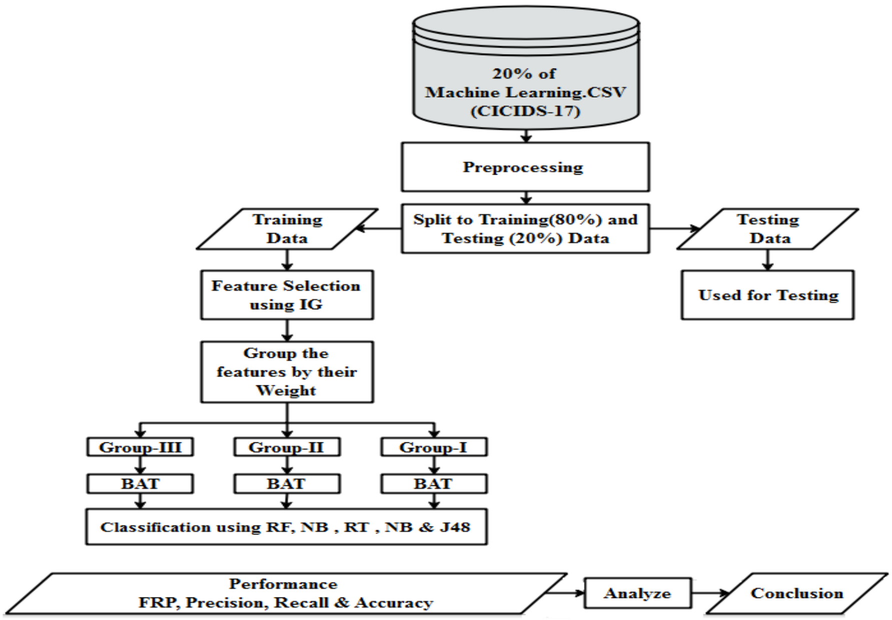
20% of
Machine Learning.CSV
(CICIDS-17)
Preprocessing
Split to Training(80%) and Testing (20%) Data
Training Data
Testing Data
Feature Selection using IG
Used for Testing
Group the features by their Weight
Group-III
Group-II
Group-I
BAT
BAT
BAT
Classification using RF, NB , RT , NB & J48
Performance
FRP, Precision, Recall & Accuracy
Analyze
Conclusion

Figure 3: The complete structure of the proposed framework

Table 4: Features retrieved from individual groups using feature selection methods

Characteristic weight	**Subset of features selected from Information gain**	**Feature subset**	**Subset of features selected from Hybrid IG-BA method**	**Feature subset**
> 0.6	15	842, 20, 54, 18, 65 years 67, 13, 12, 63, 66, 52, 40, 41, 39	8	41, 65, 8, 42, 20, 12, 66, 39
> 0.4	28	41, 13, 65, 8, 42, 20, 54, 18, 67, 12, 63, 66, 52, 40, 39, 14, 22, 36, 9, 26, 55, 24	13	13, 65,42, 54, 18, 67, 12,63, 52, 14, 22, 9, 24
> 0.3	35	41, 13, 65, 8, 42, 20, 54, 18, 67, 12, 63, 66, 52, 40, 39, 14, 22, 36, 9, 26, 55, 24, 25, 21, 2, 1, 64, 11, 16, 53, 19, 3, 37, 30, 7	21	41, 13, 42, 20, 54, 18, 67, 12, 63, 66, 52, 36, 9, 26, 55, 24, 25, 16, 37, 30
> 0.2	52	41, 13, 65, 8, 42, 20, 54, 18, 67, 12, 63, 66, 52, 40, 39, 14, 22, 36, 9, 26, 55, 24, 25, 21, 2, 1, 64, 11, 16, 53, 19, 3, 37, 30, 7, 10, 62, 28, 4, 17, 29, 5, 15, 38, 70, 27, 73, 69, 72, 31, 23, 76	34	41, 13, 42, 20, 67, 52, 40, 39, 14, 22, 36, 24, 25, 21, 64, 11, 16, 53, 19, 3, 37, 30, 7, 10, 62, 28, 4, 17, 27, 73, 69, 72, 31, 76

The results of feature selection methods is shown in Table 4, from the IG algorithm original data is grouped into subsets considering various weight threshold values 0.6,0.4,0.3, and 0.2 . The standard IG algorithms retrieved features sets of size 15, 28, 35, and 52. However, the proposed IG-BA produced feature sets of size 8, 13, 21, and 34 and are optimal features by reducing irrelevant features.

The performance of classification algorithms by applying feature set of size 15 is shown in Table 5. The Random Tree (RT) and Random Forest (RF) produced almost 95% accuracy when comparedother classification methods. However, with these features classifiers are applied to detectall attacks. Also, observed that Naïve Bayes (NB) results bad in case of the normal traffic.

Table 5: Performance of classification algorithms considering feature set of size 15

Attack/Measure	**J48**	**Random Tree(RT)**	**Bayesian Network(BN)**	**Random Forest(RF)**	**Naïve Bayes(NB)**
Normal	0.942	0.941	0.924	0.941	0.171
Attack-1	0.971	0.972	0.976	0.972	0.979
Attack-2	0.975	0.975	0.972	0.975	0.963
Attack-3	0.373	0.421	0.629	0.429	0.673
Attack-4	0.071	0.071	0.030	0.071	0.000
Attack-5	0.000	0.392	0.000	0.000	0.392
Attack-6	0.774	0.776	0.971	0.776	0.980
Recall	NA	0.951	0.943	0.946	0.885
Precision	0.946	0.946	0.934	NA	0.328
FRP	0.015	0.014	0.009	0.014	0.021

The performance of classification algorithms by applying feature set of size 28 is shown in Table 6. Random Forest (RF) produced almost 97% accuracy when compared other classification methods. The experimental results with the given classification algorithms RandomForest (RF), Random Tree (RT), and J48 are promising while detecting at Normal, Attack1 to 3. However, classification algorithms results difficulties in detecting Attack 3 and Attack 5 traffic. Moreover, it is observed that Random Tree(RT), Random Forest(RF), and J48 results lower FPR of 0.006,alsoBayesian Network(BN) results very lowest FPR i.e., 0.003. Finally it is observed that J48, Random Tree (RT), and Random Forest (RF) producedbetteraccuracy and recall of value i.e., 0.978.

Table 6: Performance of classification algorithms considering feature set of size 28

Attack/Measure	J48	Random Tree(RT)	Bayesian Network(BN)	Random Forest(RF)	Naïve Bayes(NB)
Normal	0.942	0.941	0.924	0.941	0.171
Attack-1	0.979	0.979	0.951	0.979	0.946
Attack-2	0.977	0.977	0.975	0.977	0.972
Attack-3	0.699	0.711	0.965	0.692	0.448
Attack-4	0.108	0.114	0.973	0.114	0.812
Attack-5	0.000	0.588	0.392	0.196	0.588
Attack-6	0.976	0.975	0.976	0.975	0.979
Recall	0.978	0.978	0.976	0.978	0.427
Precision	0.978	0.978	0.877	0.978	0.895
FRP	0.006	0.006	0.003	0.006	0.030

The performance of classification algorithms by applying feature set of size 35 is shown in Table 7. Random Forest (RF) produced almost 97% accuracy, recall i.e., 0.978 and a low FPR i.e., 0.004, and precision Nan when compared other classification methods. However, this classification algorithms results difficulties in detecting Attack 5 traffic.The experimental results with the given classification algorithms Random Forest (RF), Random Tree (RT), and J48 are promising while detecting at Attack1 to 3 and produced better FRP. Finally it is observed that Naïve Bayes(NB) produce low FRP.

Table 7: Performance of classification algorithms considering feature set of size 35

Attack/Measure	J48	Random Tree(RT)	Bayesian Network(BN)	Random Forest(RF)	Naïve Bayes(NB)
Normal	0.969	0.969	0.899	0.969	0.347
Attack-1	0.969	0.967	0.952	0.969	0.701
Attack-2	0.969	0.964	0.962	0.966	0.961
Attack-3	0.677	0.754	0.959	0.739	0.553
Attack-4	0.126	0.721	0.956	0.764	0.821
Attack-5	0.000	0.388	0.582	0.000	0.776
Attack-6	0.965	0.966	0.964	0.967	0.954
Recall	0.969	0.968	0.920	0.979	0.434
Precision	NaN	0.968	0.965	NaN	0.897
FRP	0.006	0.005	0.006	0.004	0.018

Similarly, while considering 52 features Random Forest (RF) produced accuracy of 97.8%, recall i.e., 0.979, and FPR i.e., 0.004 compared to other classification algorithms. However, the precision recordedNaN. From this it is noted that this algorithms failed to detect Attack 5.

6. Conclusions

The proposed method validates that feature selection improves the performance of feature selection on anomaly detection data. The proposed feature selection produces the ranking of features based on their weight values using IG algorithm, resulting in a subset of features to rank. Later, individual subset applied to BA algorithms and then processed which results optimal features for the further classification. From the overall Random Forest performs promising using all sizes of feature sets from 15, 28,35, and 52. Also noticed that J48 results better in case of featuresets of 35 and 52. All the traffics detects properly using feature subsets of 35, and52. However, the Bayes Naïve (BN) results low accuracy compared other classifiers. Also notice in this classification subset of features impact on reduction of FPR.In the future, work plan to conduct study on multi classification.

References

1. K. Scarfone, P. Mell, Special Publication 800-94: Guide to Intrusion Detection and Prevention Systems (IDPS), National Institute of Standards and Technology (NIST) (2007).
2. Bace, Rebecca Gurley: Intrusion detection. Copyright 2000 by Macmillan Technical Publishing, ISBN 1-57870-185-6.
3. J. Zhang, H. Li, Q. Gao, H. Wang and Y. Luo, “Detecting Anomalies from Large Network Traffic Data Using an Adaptive Detection Approach,” Inf. Sci., Vol. 318, pp. 91-110, October 2015.
4. García-Teodoro, P .; Díaz-Verdejo, J .; Maciá-Fernández, G .; Vázquez, E. Anomaly-based network intrusion detection: techniques, systems, and challenges. Comput. Secur. 2009, 28, 18–28. [CrossRef]
5. R. Panigrahi and S. Borah, `` A Detailed Analysis of the CICIDS2017 Dataset for the Design of Intrusion Detection Systems ", Int. J. Eng. Technol., Vol. 7, no. 24, pp. 479–482, 2018
6. Alazab, A .; Hobbs, M .; Abawajy, J .; Alazab, M. Using the feature selection for the intrusion detection system. In Proceedings of the 2012 International Symposium on Communications and Information Technologies (ISCIT), Gold Cost, Australia, October 2-5, 2012; 296-301.
7. MK Kundu, DP Mohapatra, A. Konar, and A. Chakraborty, `` Decision Tree Techniques Applied to NSL-KDD Data and Its Comparison with Various Feature Selection Techniques, " Smart Innov. Syst. Technol., Vol. 27, no. 1, pp. 205-211, 2014.
8. W. Wang, Y. He, J. Liu, and S. Gombault, “Building Important Features from Massive Network Traffic for Light Intrusion Detection,” IET Inf. Secur., Vol. 9, no. 6, pp. 374–379, November 2015
9. I. Ahmad, M. Hussain, A. Alghamdi and A. Alelaiwi, “Improving SVM Performance in Intrusion Detection Using the Selection of Optimal Feature Subsets Based on Key Genetic Components”, Neural Comput. Appl., Vol. 24, nos. 7–8, pp. 1671-1682, June 2014.
10. S.-H. Kang and KJ Kim, “A feature selection approach to find optimal feature subsets for the network intrusion detection system”, Cluster Comput., Vol. 19, no. 1, pp. 325–333, March 2016.
11. AI Madbouly, SA King Abdulaziz University Jeddah, AM Gody and TM Barakat, `` Relevant feature

selection model using data mining for intrusion detection system ", Int. J. Eng. Trends Technol., Vol. 9, no. 10, p. 501-512, March 2014.

12. E. Popoola and A. Adewumi, “Efficient feature selection technique for a network intrusion detection system using discrete differential evolution and a decision tree”, Int. J. Netw. Secur., Vol. 19, no. 5, pp. 660–669, 2017.
13. BA Tama and KH Rhee, “A Combination of PSO-Based Feature Selection and Tree-Based Classifier Set for Intrusion Detection Systems”, Adv. Comput. Sci. Ubiquitous Comput., Vol. 373, pp. 489–495, February 2015.
14. A. Yulianto, P. Sukarno and N. Suwastika, `` Improving the performance of the AdaBoost-based intrusion detection system (IDS) on the CIC IDS 2017 dataset ", J. Phys., Conf. Ser., Vol. 1192, March 2019, art. no. 012018, doi: 10.1088 / 1742-6596 / 1192/1/012018.
15. S. Bhattacharya and S. Selvakumar, “Multi-measure multi-weight ranking approach for the identification of network characteristics for the detection of DoS and probe attacks”, Comput. J., vol. 59, no. 6, p. 923–943, June 2016.
16. TA Alhaj, MM Siraj, A. Zainal, HT Elshoush and F. Elhaj, “Feature Selection Using Information Gain for Better Structure-Based Alert Correlation,” PLoS ONE, vol. 11, no. 11, 2016, art. no. e0166017.
17. Guyon, I., Elisseeff, A. (2003). An introduction to selecting variables and features. Journal of machine learning research 3 (March) 1157-1182.
18. Saeys, Y., Inza, I., Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics 23 (19) 2507-2517.
19. Yang, XS. (2010). A new algorithm inspired by metaheuristic bats. In: Nature Inspired Cooperative Strategies for Optimization (NICSO 2010) (Eds.Cruz C., Gonzalez J., Krasnogor N. and Terraza G.), Springer, SCI 284, pp 65-74.
20. Yang, XS. (2008). Metaheuristic algorithms inspired by nature. Luniver Press.
21. Xin-She Yang and Amir H. Gandomi. 2012. Bat Algorithm: A New Approach for Global Engineering Optimization. Engineering Computations, 29 (5), 464-483.
22. M. Reazul, A. Rahman and T. Samad, “A Network Intrusion Detection Framework Based on a Bayesian Network Using a Wrapper Approach”, Int. J. Comput. Appl., Vol. 166, no. 4, p. 13-17, May 2017.
23. J. Jiang, Q. Wang, Z. Shi, B. Lv and B. Qi, `` RST-RF: A hybrid model based on approximate set theory and a random forest for network intrusion detection " , in Proc. ACM Int. Conf. Process., 2018, p. 77-81.
24. RK Singh, S. Dalal, VK Chauhan and D. Kumar, “Optimizing FAR in an Intrusion Detection System Using a Random Forest Algorithm”, SSRN Electron. J., vol. 5, pp. 3-6, March 2019.
25. N. Ding, H. Gao, H. Bu and H. Ma, `` RADM: Real-time Anomaly Detection in Multivariate Time Series Based on a Bayesian Network ", in Proc. IEEE Int. Conf. Smart Internet Things, August 2018, p. 129-134.
26. K. Goeschel, `` Reducing False Positives in Intrusion Detection Systems Using Data Mining Techniques Using Supporting Vector Machines, Decision Trees, and Naive Bayes for Offline Analysis ", in Proc. SoutheastCon, March 2016, p. 1–6.
27. S. Shakya and S. Sigdel, `` An Approach to Developing a Hybrid Algorithm Based on a Support Vector Machine and Naive Bayes for Anomaly Detection ", in Proc. Int. Conf. Comput. Common. Autom. (ICCCA), January 2017, pp. 323–327.
28. R. Chitrakar and H. Chuanhe, `` Anomaly Detection Using Classification of Support Vector Machines with Clustering k-Medoids ", in Proc. 3rd Asian Himalayas Int. Conf. Internet, November 2012, p. 1–5, doi: 10.1109 / AHICI. 2012.6408446.
29. AP Muniyandi, R. Rajeswari and R. Rajaram, “Network Anomaly Detection by Cascading K-means Clustering and C4.5 Decision Tree Algorithm”, Procedia Eng., Vol. 30, pp. 174-182, February 2012.
30. [41] S. Thaseen, intrusion detection model using the fusion of PCA and optimized SVM. Boca Raton, FL, USA: CRC Press, 2014, pp. 879–884.