International Journal of Engineering & Technology, 7 (4) (2018) 5934-5939

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.22844
Research paper

The precocious classification framework for network intrusion detection system

D. Selvamani [1] *, V. Selvi [2]

[1] Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamil Nadu
[2] Assistant Professor, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamil Nadu
**Corresponding author E-mail: *selvamani.bhaskar@gmail.com*

Abstract

The Intrusion Detection System (IDS) can be used broadly for securing the network. Intrusion detection systems (IDS) are typically positioned laterally through former protecting safety automation, like access control and verification, as a subsequent line of resistance that guards data classifications. This projected framework established on the precocious feature selection, which is consent to lessens the number of features generated in the KDD CUP 99 benchmark dataset. The projected framework customs the Back Propagation Neural Networks to recognize the Denial of Service (DoS), where it is a combined variety of attack in the networks.

Keywords: *Intrusion Detection System; KDD CUP 99; Feature Selection; Information Gain; Artificial Neural Network*

1. Introduction

Internet exhibits a demanding part in this moderate world. It is practiced in the shopping [14], education [15], social networking [16], business [17], etc. It has enhanced a risk of computer classifications associated to the internet fetching the objectives of intrusions by cybercriminals [1]. Cyber criminal's outbreak the arrangements to advance illegal access to data, exploit data or to lessen the obtainability of information to the legal users. As a result, massive financial losses to companies and lose their kindness to the customer. Intrusion avoidance techniques such as user authentication (e.g., using biometrics or password), information protection (e.g., Encryption), sidestep programming errors and firewalls have been practiced to protect computer systems [18-21]. Principally, IDS design [R] and its instigation can be either network based (NIDS) or host Based (HIDS). NIDS is an intrusion recognition system that apprehensions data packets roving on the network and similar them to a database of signatures. Today, most of the accessible IDS tools are sensing unsolicited, i.e. malevolent activities or movements by assessing TCP/IP Connections or Log files, for attacks in an instance. These IDS systems are employed to attack an spell prepared on the network with the numerous innovative fears in a network.

2. Related works

Elbasiony, Reda M., et al. [3] The present research anticipated a mixture recognition framework that is determined by on data mining arrangement and collecting methods. In harm recognition, random forests classification algorithm is deployed to form intrusion forms mechanically from a training dataset. In irregularity recognition, the k-means collecting algorithm is practiced to sense new intrusions by clustering the network associates.
Ahmad, Iftikhar, et al. [4] projected a new-fangled intrusion detection structure based on *K*-nearest neighbor (*K*-nearest neighbor, referred to as KNN below) classification algorithm in a wireless sensor network. This structure can isolate anomalous nodes from regular nodes by discerning their anomalous behaviors, and the authors investigated parameter selection and inaccuracy rate of the intrusion detection system.
Aburomman, Abdulla Amin, and Mamun Bin IbneReaz [5] proposed a novel collaborative structure technique that employs PSO engendered weights to produce collaborative of classifiers with improved precision for intrusion detection. Local uni-modal sampling (LUS) technique is employed as a meta-optimizer to discover healthier behavioral constraints for PSO. Besides, the current research exploited the KDD CUP 99 dataset for discovering the intrusion detection.
Thaseen, Ikram Sumaiya in [6] proposed an intrusion detection model consuming chi-square feature selection and multi-class support vector machine (SVM). In this paper, a constraint tuning system is implemented for optimization of Radial Basis Function kernel parameter. The NSL-KDD dataset which is an enriched version of KDDCup 1999 dataset was practiced in this paper.
Li, Longjie, et al [7] a unique hybrid model was proposed with the determination of identifying network intrusion meritoriously. In the proposed model, Gini index is accustomed to picking the most excellent division of features, the gradient based decision tree (GBDT) algorithm is implemented to sense network attacks. The particle swarm optimization (PSO) algorithm is exploited to augment the parameters of GBDT. The NSL-KDD dataset was used to assess the activity of the future technique.

3. Problem identification

It is highly acknowledged that anomaly-based IDS writhe from the significant rate of deceitful alarms. It is to consider that intrusion detection is a data study method and can be premeditated as a tricky of categorizing data accurately. From this standpoint, it can also be detected that several classification schemes is as decent as the data accessible to it as input. Further, clean the data, excep-

tional perfect results are probable to be accomplished. From anomaly-based IDS point of view, it entails that if we can abstract the features that delineate characteristic data from atypical one appropriately, a false positive rate can be condensed to an excessive level. We detect that most of the data mining and machine learning based techniques in intrusion detection make a practice of essential tools and methods. It may turn out that these universal methods are not very operational in categorizing data as normal or abnormal with very high precision.

4. Proposed precocious classification framework

In this proposed innovative classification framework has the three phases i) Pre-processing ii) Feature Selection stage, iii) Classification stage [22, 23] are integrated to acquire the perfect results. Figure 1 portrays the projected Precocious classification framework for Intrusion Detection System.

Phase 1: Pre-Processing: In this phase, Standardization is exploited to competent the classification exactness in the sorting stage.

Phase 2: Feature Selection: In this phase, a filter feature selection method and Classifier are mongrelized to acquire the durable significant features with determined classification accuracy and least error rates. The algorithm suggested in this phase is created as the Precocious feature selection method. This algorithm picks the features established on the exactness by the classification technique ANN.

Phase 3: Classification Stage: In this phase, Artificial Neural Network has employed to categorize the information into three classes Secure, Known attacks and Unknown attacks.

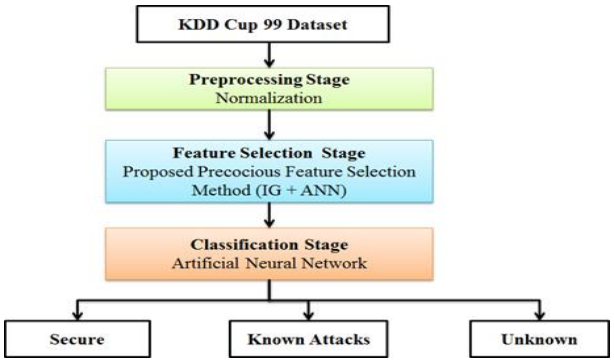

Fig. 1: Proposed Precocious Classification Framework for Network-Based Intrusion Detection.

4.1. Pre-processing by using normalization

According to the following equation, each numerical is value set between 0.0 and 1.0. ANN training is frequently more competent with normalized data; it is accustomed as the superior predictor.

$$a = \frac{a-min}{max-min}$$

Where x is the normalized value with a series between 1 and 0, a is the unique value, min and max are minimum and maximum values of the original variable. These principles are exploited in corresponding to the upper and lower limits of the stimulation task sigmoid which are practiced in the ANN models.

4.2. Information gain feature selection method

In this proposed articulate feature selection method, the data advance the feature selection method and Artificial Neural Network classification method has hybridized to select the powerful relevant and low redundant features since the KDD CUP 99 dataset.

Entropy is frequently used in the information theory measure, which exemplifies the transparency of a random collection of samples [8]. It is in the establishment of Gain Ratio, Information Gain and Similarity Uncertainty (SU) [9]. Therefore, the entropy quantity is measured as a parameter of the classification's randomness. The entropy of B is

$$H(B) = \sum_{b\in Y} p(b)\log_2(p(b)) \quad (1)$$

Where p(b) is the marginal probability density function for the arbitrary variable B. If the experimental values of B in the training data set S are segregated in bestowing to the values of a second feature A, and the entropy of B in reference to the segregations persuaded by A is less than the entropy of B prior to segregating, at that point there is an association between features B and A. The entropy of B after spotting A is then:

$$H(B|A) = \sum_{a\in A} p(a)\sum_{b\in B} p(b|a)\log_2\big(p(b|a)\big) \quad (2)$$

Where p (b |a) is the conditional probability of b given a.

As given the entropy is a measure for contamination in a training set S, we can state a measure replicating supplementary data nearly B provided by A that epitomizes the quantity which the entropy of B decreases. This amount is known as IG. It is known by

$$IG = H(B) - H(B|A) = H(A) - H(A|B) \quad (3)$$

IG [8] is a proportioned measure, and it is known by equation (3). The information gained about B after observing A is alike the information gained approximately A after detecting B. A flaw of the IG criterion is that it is subjective in accord of features with further values even once they are not highly instructive.

4.3. Proposed precocious feature selection method (PFSM)

The various filter feature selection algorithms established on information theory [10] have been projected for high-dimensional data sets; meanwhile, they are profligate and competent. On the other hand, once there is a lesser quantity of samples, the probable multivariate data befits erroneous, and the classification enactment of these algorithms worsened. From this time, the wrapper methods are more appropriate, although they are more arduous. The proposed PFSM algorithm for selecting features established on their interaction information. PFSM picks a feature if totaling it to the presently nominated subset enhances the precision considerably (wrapper approach), however, it quests more competently by figuring the interaction information (filter approach) to discover candidate features to be added to the existing subset. Furthermore, the PFSM algorithm deploys prompt ending to avert overfitting.

Stage 1: Initialization: PFSM initiates through an empty set E and picks the key feature a_k from the full set FS of N features that stretches the leading mutual information between the feature and the class target CT.

$$a_k = \arg\max_{a_j \in FS}[IG(a_j, CT)]$$

The feature a_k is further to the nominated feature set E (i.e., $E \leftarrow E \cup a_k$) and it is detached from the set FS (i.e., $FS \leftarrow FS \setminus a_d$)

Stage 2: Finding a candidate feature: A good candidate feature ought to capitalize on the interaction information and class relevance to the present subset E. Then, the next candidate feature a_d is to be added to the subset S is the one that makes the most of the Joined mutual Information criterion [20]:

$$a_d = \arg\max_{a_j \in FS}[IG(a_j;\ CT) + \frac{1}{|E|}\sum_{a_i\in E} IG(a_j; a_i; CT)]$$

The above equation is the relevance term that processes the information gain with a_j, whereas the second term in the above equation of the interaction information between a_j, a_i (currently selected features), and CT. As a result, the JMI criterion opts the candi-

date feature a_d that is pertinent to the target class and has durable interaction with those that are previously designated.

Stage 3: Conditional Inclusion: The candidate feature a_d provisionally added to the current set E (i.e E∪ a_d). Here we calculate the classification accuracy for the training set with the following subset exhausting as a particular classifier. If the classification accuracy for the training set expands considerably, go to Stage 4. Otherwise, a_d is not selected and then detached from the set FS. If FS is empty, dismiss the algorithm. If not, go to Stage 2.

A k-fold cross-validation is employed to assess the classification exactness of a specific classifier for a set of features. Student's paired right-tailed t-test (at 0.1 level) is directed to calculate if the classification accuracy of the training set by means of the candidate subset S∪ x_d is ominously better (>>) than the using the subset E. The confidence level is opted at 0.1 to consent more features to be added, as noted in [14].

Stage 4: Incremental Inclusion: We calculate the classification accuracy for the validation set through the subset E ∪ a_d by a given classifier. If the classification accuracy for the validation set does not fall expressively (using a Student's paired left-tailed t-test at 0.1 level), perpetually add a_d into the set E (i.e., E ←E∪ a_d) and to eliminate a_d from the set FS, to apprise the classification accuracy rates for the training and validation sets, and go to Stage 2. Then, a_d is not nominated, and the algorithm dismisses.

Step by Step Procedure for proposed Precocious feature Selection algorithm

Input: A sample set S with a full feature set FS of N features, a target class CT, and a given ANN classifier.

Output: The selected feature set FS

Step 1: Utilize the k-fold cross-validation on S to create S_{train} and S_{val}

Step 2: for every feature $a_i \in FS$

Step 3:InfoGain(i) = IG(a_i, CT)

Step 4: End for

Step 5:$a_k = \arg\max_{a_i \in FS}(\text{InfoGain(i)})$

Step 6:$BAcc_{train} = \text{Classifier}\ (S_{train}, S_{train}, a_k)$

Step 7:$BAcc_{val} = \text{Classifier}\ (S_{train}, S_{val}, a_k)$

Step 8:$E = \{a_k\}$

Step 9:$FS = FS \backslash a_k$

Step 10: While $FS \neq \emptyset$ (empty set)

Step 11: for each feature $a_j \in FS$

Step 12: Interaction (j) = $IG(a_j;\ CT) + \frac{1}{|E|}\sum_{a_i \in E} IG(a_j; a_i; CT)$

Step 13: end for

Step 14:$a_d = \arg\max_{a_j \in FS}(\text{Interaction(j)})$

Step 15:$FS = FS \backslash a_d$

Step 16:$E_{tmp} = E\ \cup \{a_d\}$

Step 17:$Acc_{train} = \text{Classifier}\ (S_{train}, S_{train}, E_{tmp})$

Step 18:if ($Acc_{train} \gg BAcc_{train}$)

Step 19:$Acc_{train} = \text{Classifier}\ (S_{train}, S_{val}, E_{tmp})$

Step 20:if ($Acc_{val} \ll BAcc_{val}$)

Step 21: break

Step 22: else

Step 23:$BAcc_{train} = Acc_{train}$

Step 24:$BAcc_{val} = Acc_{val}$

Step 25: $E = E_{tmp}$

Step 26: end if

Step 27: end if

Step 28: end while

4.4. Proposed precocious feature selection method (PFSM)

The proposed Precocious Classification Framework (PCF) practices a renowned supervised learning of neural network architecture [11] recognized as Multi-Layer-Perceptron (MLP) with back-propagation gradient-descent in the premeditated PCF. To formulate a feed-forward multi-layer in MLP, the pool of non-linear neurons is associated with one another. This technique is identified to be very advantageous for classification and prediction disputes. Cross-validation is deployed to regulate the 'optimal' number of hidden layers and neurons which depended on the investigational scheme of the IDS. In particular, the training of the MLP instigated from a less number of neurons, in addition only one hidden layer, which processes the error ratio of the trained BP on holdout samples, gradually accumulative the number of neurons at the hidden layer in which the recital of the accomplished stage on holdout samples has originated to go down as a result of the problem of overtraining. Accordingly, we acquire the best number of neurons for the hidden layer of the ANN [12].

5. Result and discussion

5.1. Description of the dataset

Aware of the deficiency of appropriate audit data sets for intrusion detection, KDDCUP [13] sets out (1) to create an intrusion-detection assessment corpus which could be assumed by various scholars, (2) to evaluate numerous intrusion-detection systems, (3) to initiate a ample form of attacks and (4) to compute both attack-both false-alarm rates and recognition rates for accurate and consistent traffic. The subsequent are the varieties of attacks in 1999 KDD CUP dataset. (a) Disowning of Service, (b) Examining, (c) Remote to Local (R2L), and (d) User to Root (U2R). The subsequent table 1 stretches the depiction of the dataset.

Table 1: Description of the KDD CUP 99 Dataset

Sl.No	Feature Name	Description
1	Duration	length (number of seconds) of the connection
2	Protocol_type	type of protocol (e.g., TCP, UDP, etc.)
3	Service	network service on the destination, e.g., HTTP, telnet, etc.
4	Src_bytes	the quantity of data bytes from source to destination
5	Dst_bytes	number of data bytes from destination to source
6	Flag	normal or error status of the connection
7	Land	1 if the connection is from/to the same host/port; 0 otherwise.
8	Wrong_fragment	number of 'wrong' fragments
9	Urgent	number of urgent packets
10	hot	Number of 'hot'' indicators
11	Num_failed_logins	Number of failed login attempts
12	Logged_in	1 if successfully logged in ; 0 otherwise
13	Num_compromised	Number of 'compromised' conditions
14	Root_shell	1 if root shell has reached; 0 otherwise
15	Su_attempted	1 if 'su root' command attempted; 0 otherwise
16	Num_root	Number of 'root' accesses
17	Num_file_creations	Number of file creation operations
18	Num_shells	Number of shell prompts
19	Num_access_files	Number of operations on access control files
20	Num_outbound_cmds	The quantity of outbound commands in an FTP session
21	Is_hot_login	1 if the login belongs to the 'hot' list; 0 otherwise

22	Is_guest_login	1 if the login is a 'guest' login ; 0 otherwise
23	count	number of connections to the same host as the current connection in the past two seconds
24	serror_rate	% of connections that have ``SYN" errors
25	rerror_rate	% of connections that have ``REJ" errors
26	same_srv_rate	% of connections to the same service
27	diff_srv_rate	% of connections to different services
28	srv_count	the number of connections to the identical service as the current connection in the past two seconds
29	srv_serror_rate	% of connections that have 'SYN' errors
30	srv_rerror_rate	% of connections that have 'REJ' errors
31	srv_diff_host_rate	% of connections to different hosts
32	dst_host_count	No. of connections to the same host as the current connection in the past two seconds
33	dst_host_serror_rate	% of connections that have 'SYN' errors
34	dst_host_rerror_rate	% of connections that have 'REJ' errors
35	dst_host_same_srv_rate	% of connections to the same service
36	dst_host_diff_srv_rate	% of connections to the different services
37	dst_host_srv_count	No. of connections to the same service as the current connection in the past two seconds
38	dst_host_srv_serror_rate	% of the connections that have "SYN" errors
39	dst_host_srv_rerror_rate	% of the connections that have "REJ" err= 1ors
40	dst_host_srv_diff_host_rate	% of the connections to different hosts
41	dst_host_sam_src_port_rate	% of the connections to destination with same source port value
42	dst_host_diff_src_port_rate	% of the connections to the destination with different source port value
43	Class	1-yes or 0-No

5.2. Result obtained by PFSM

Following table 2 provides the outcome attained by the proposed Precocious Feature Selection method and current filter-based feature selection techniques like Gain Ratio, Symmetrical Uncertainty, and Information Gain. From table 2, Gain Ratio filters 32 features, Symmetrical Uncertainty filters 31 features, Information Gain screens only 27 features, and the proposed PFSM gives only 23 features. To assess the competence of the proposed PFSM and other approaches by consuming classification techniques like Artificial Neural Network (ANN), Support Vector Machine. The assessment of metrics is like Accuracy, Error rates, True Positive Rate, False Positive Rate, Precision, Recall and ROC curves.

Table 2: Number of Features Obtained by Existing Feature Selection Methods and Proposed Precocious Feature Selection Method

Gain Ratio	Symmetrical Uncertainty	Information Gain	Proposed Precocious Feature Selection Method
is_guest_32	src_bytes	Src_bytes	dst_host_srv_count
logged_in	dst_bytes	dst_host_same_srv_rate	dst_host_same_srv_rate
Dst_bytes	dst_host_srv_count	Dst_bytes	Src_bytes
dst_host_srv_serror_rate	dst_host_same_srv_rate	dst_host_rerror_rate	dst_host_rerror_rate
Src_bytes	dst_host_rerror_rate	Service	dst_host_serror_rate
dst_host_serror_rate	Service	dst_host_diff_srv_rate	Dst_bytes
dst_host_same_srv_rate	dst_host_srv_rerror_rate	dst_host_srv_rerror_rate	dst_host_srv_rerror_rate
dst_host_srv_count	dst_host_serror_rate	srv_Count	dst_host_srv_serror_rate
dst_host_srv_rerror_rate	logged_in	count	dst_host_count
dst_host_rerror_rate	dst_host_srv_serror_rate	dst_host_serror_rate	srv_Count
Service	dst_host_diff_srv_rate	dst_host_srv_serror_rate	dst_host_diff_srv_rate
dst_host_count	srv_Count	dst_host_same_src_port_rate	count
hot	dst_host_count	logged_in	Service
dst_host_diff_srv_rate	dst_host_same_src_port_rate	dst_host_count	dst_host_same_src_port_rate
srv_rerror_rate	count	rerror_rate	srv_rerror_rate
dst_host_same_src_port_rate	flag	srv_rerror_rate	serror_rate
Flag	srv_rerror_rate	same_srv_rate	srv_serror_rate
srv_count	rerror_rate	diff_srv_rate	dst_host_srv_diff_host_rate
rerror_count	is_guest_32	dst_host_srv_diff_host_rate	rerror_rate
dst_host_srv_diff_host_rate	dst_host_srv_diff_host_rate	serror_rate	same_srv_rate
srv_serror_rate	same_srv_rate	srv_serror_rate	diff_srv_rate
count	diff_srv_rate	is_guest_32	srv_diff_host_rate
serror_rate	hot	hot	Protocol_type
diff_srv_rate	srv_serror_rate	Protocol_type	
Protocol_type	hot	srv_diff_host_rate	
same_srv_rate	srv_serror_rate	num_compromised	
srv_diff_host_rate	serror_rate	num_failed_32s	
num_compromised	Protocol_type		
num_failed_32s	srv_diff_host_rate		
Wrong_fragment	num_compromised		
land	num_failed_32s		
urgent			

The following figure 2 represents the number of features in the unique dataset, and the number of features attained by present feature selection approaches like Information Gain, Gain Ratio, Symmetrical Uncertainty and proposed Precocious Feature Selection method.

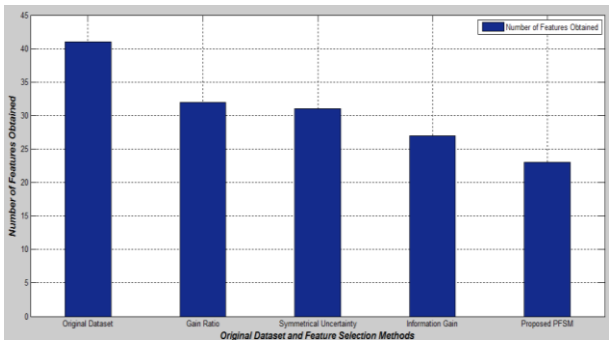

Fig. 2: Number of Features Obtained Gain Ratio, Symmetrical Uncertainty, Information Gain and Proposed Precocious Feature Selection Method.

Following table 3 gives the performance analysis of the original dataset, Information Gain and proposed PFSM using Artificial Neural Network as the classifier. From table 3 it is strong that the projected technique performs effectively and contributes the maximized accuracy, Kappa Statistics, TPR, FPR, Precision, Recall, F-Measure, and ROC Area.

Table 3: Performance Analysis of the Original Dataset, Information Gain and Proposed Precocious Feature Selection Using Artificial Neural Network Classification Method

Evaluation Metrics	Original Dataset	Information Gain	Proposed Precocious Feature Selection Method
Accuracy	69.3333 %	93.7677 %	98.5 %
Relative Absolute Error (RRAE)	45.3867 %	43.384 %	41.162 %
Root Relative Squared Error (RRSE)	88.8892 %	77.864 %	59.1282 %
Kappa Statistics	0.5539	0.6242	0.7683
True Positive Rate	0.682	0.938	0.982
False Positive Rate	0.16	0.283	0.364
Precision	0.642	0.934	0.983
Recall	0.682	0.938	0.982
F-Measure	0.612	0.942	0.987
ROC Area	0.836	0.868	1

Following table 4 gives the performance analysis of the original dataset, Information Gain and proposed PFSM using Support Vector Machine as the classifier. From table 4 it is clear that the proposed method performs well and gives the maximized accuracy, Kappa Statistics, TPR, FPR, Precision, Recall, F-Measure, and ROC Area.

Table 4: Performance Analysis of the Original Dataset, Information Gain and Proposed Precocious Feature Selection Using Support Vector Machine Classification Method

Evaluation Metrics	Original Dataset	Information Gain	Proposed Precocious Feature Selection Method
Accuracy	66.4312 %	75.3333 %	91.5755 %
Relative Absolute Error (RRAE)	92.8816 %	97.2935 %	97.2935 %
Root Relative Squared Error (RRSE)	93.562 %	88.2774 %	63.2201 %
Kappa Statistics	0.3553	0.4118	0.6543
True Positive Rate	0.642	0.742	0.916
False Positive Rate	0.22	0.26	0.251
Precision	0.494	0.647	0.865
Recall	0.642	0.742	0.916
F-Measure	0.545	0.691	0.691
ROC Area	0.8	0.822	0.943

Figure 3a, 3b and 3c characterizes the presentation analysis of the classification methods like Artificial Neural Network and Support Vector Machine for the particular original dataset; features attained by Information Gain and proposed Precocious feature selection method. Moreover, the statistics are to assess the efficiency of the original dataset, information gain and proposed PFSM. From the given figures it is vibrant that the proposed Precocious feature selection method accomplishes productively in all features when ANN as the classifier for the intrusion detection system.

Figure 3a elucidates the Performance analysis on the classification accuracy of the classification Methods like Artificial Neural Network, Support Vector Machine for the given original dataset, Information gain processed dataset and Proposed PFSM processed dataset. (X-Axis give the classification methods, Y-Axis epitomizes the classification accuracy in % for original dataset, Information Gain and Proposed PFSM).

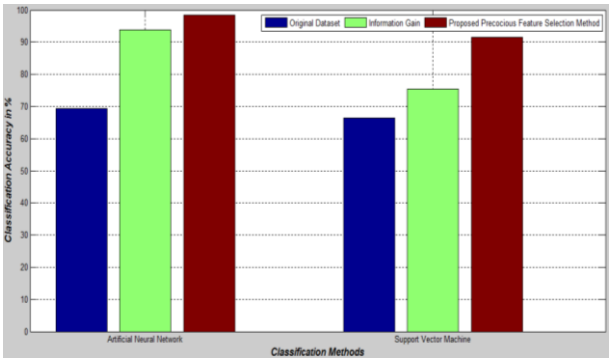

Fig. 3: A) Performance Analysis on the Classification Accuracy of the Classification Methods Like Artificial Neural Network, Support Vector Machine for the Given Original Dataset, Information Gain Processed Dataset and Proposed PFSM Processed Dataset.

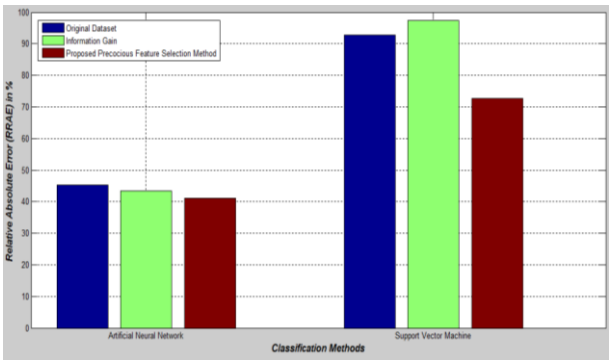

Fig. 3b: Shows the Performance Analysis on Relative Absolute Error (RAE) of the Classification Methods Like Artificial Neural Network, Support Vector Machine for the Given Original Dataset, Information Gain Processed Dataset and Proposed PFSM Processed Dataset. (X-Axis Give the Classification Methods Y-Axis Symbolizes the Relative Absolute Error (RAE) In % for Original Dataset, Information Gain and Proposed PFSM).

Figure 3c describes the Performance analysis on Root Relative Squared Error (RRSE) of the classification Methods like Artificial Neural Network, Support Vector Machine for the given original dataset, Information gain processed dataset and Proposed PFSM processed dataset. (X-Axis give the classification methods, Y-Axis signifies the Root Relative Squared Error (RRSE) in % for original dataset, Information Gain and Proposed PFSM).

Figure 3b: Performance analysis on Relative Absolute Error (RAE) of the classification Methods like Artificial Neural Network, Support Vector Machine for the given original dataset, Information gain processed dataset and Proposed PFSM processed dataset

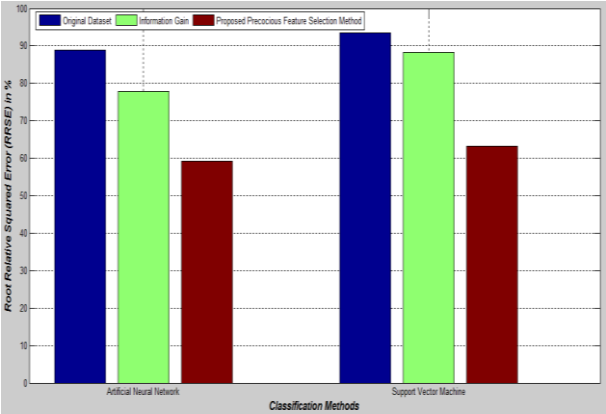

Fig. 3c: Performance Analysis on Root Relative Squared Error (RRSE) of the Classification Methods Like Artificial Neural Network, Support Vector Machine for the Given Original Dataset, Information Gain Processed Dataset and Proposed PFSM Processed A Dataset

6. Conclusion

The pre-processing technique has practiced confiscating the redundant and irrelevant features from the dataset. This approach has exploited to augment the prediction accuracy. In this contribution, an innovative Precocious Classification Framework has planned to heighten the accuracy of the classification in the IDS. The Proposed PFSM method has led to confiscate the irrelevant feature in the IDS dataset for the classification of the network. From the outcomes attained it has been substantiated that the proposed methodology accomplished healthier than the common feature selection technique in the Intrusion Detection System. Moreover, also it advances the prediction accuracy and lessens the error rates. The minimization of error rates outcomes the excellent classification accuracy.

References

[1] Shengyi Pan, Thomas Morris and Uttam Adhikari, "Developing a hybrid intrusion detection system using data mining for power systems," *IEEE Transactions on Smart Grid*, Vol.6, No.6, pp.3104-3113, 2015. https://doi.org/10.1109/TSG.2015.2409775.

[2] Syed Ali Raza Shah and Biju Issac, "Performance comparison of intrusion detection systems and application of machine learning to Snort system," *Future Generation Computer Systems*, Vol. 80, pp.157-170, 2018. https://doi.org/10.1016/j.future.2017.10.016.

[3] M. Elbasiony, Reda, *et al*, "A hybrid network intrusion detection framework based on random forests and weighted k-means," *Ain Shams Engineering Journal*, Vol. 4, No. 4, pp.753-762, 2013. https://doi.org/10.1016/j.asej.2013.01.003.

[4] Iftikhar Ahmad, *et al., "*Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components," *Neural computing and applications*, Vol.24, No.7-8, pp.1671-1682, 2014. https://doi.org/10.1007/s00521-013-1370-6.

[5] Aburomman, Abdulla Amin and Mamun Bin IbneReaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," *AppliedSoftComputing*, Vol. 38, pp.360-372, 2016. https://doi.org/10.1016/j.asoc.2015.10.011.

[6] Thaseen, Ikram Sumaiya and CherukuriAswani Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *Journal of King Saud University-Computer and Information Sciences*, Vol. 29, No.4, pp.462-472, 2017. https://doi.org/10.1016/j.jksuci.2015.12.004.

[7] Li, Longjie, et al., "Towards Effective Network Intrusion Detection: A Hybrid Model Integrating Gini Index and GBDT with PSO," Journal of Sensors, 2018. https://doi.org/10.1155/2018/1578314.

[8] Jadhav, Swati, Hongmei He, and Karl Jenkins. "Information gain directed genetic algorithm wrapper feature selection for credit rating." *Applied Soft Computing* 69 (2018): 541-553. https://doi.org/10.1016/j.asoc.2018.04.033.

[9] Venkataraman, Sivakumar, and Rajalakshmi Selvaraj. "Optimal and Novel Hybrid Feature Selection Framework for Effective Data Classification." *Advances in Systems, Control and Automation*. Springer, Singapore, 2018. 499-514. https://doi.org/10.1007/978-981-10-4762-6_48.

[10] Ramírez-Gallego, Sergio, et al. "An information theory-based feature selection framework for big data under apache spark." *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48.9 (2018): 1441-1453. https://doi.org/10.1109/TSMC.2017.2670926.

[11] MR Gauthama Raman, *et al.*, "A hypergraph and arithmetic residue-based probabilistic neural network for classification in intrusion detection systems," *NeuralNetworks*, Vol. 92, pp.89-97, 2017. https://doi.org/10.1016/j.neunet.2017.01.012.

[12] Abien Fred M. Agarap, "A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data," *Proceedings of the 2018 10th International Conference on Machine Learning and Computing, ACM*, 2018.

[13] Dataset Source: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[14] Lakshmanaprabu SK, K. Shankar, Deepak Gupta, Ashish Khanna, Joel J. P. C. Rodrigues, Plácido R. Pinheiro, Victor Hugo C. de Albuquerque. "Ranking Analysis for Online Customer Reviews of Products Using Opinion Mining with Clustering". Complexity, 2018: 1-9. https://doi.org/10.1155/2018/3569351.

[15] Eka Sugiyarti, Kamarul Azmi Jasmi, Bushrah Basiron, Miftachul Huda, K. Shankar, Andino Maseleno, "Decision Support System of Scholarship Grantee Selection using Data Mining", International Journal of Pure and Applied Mathematics, 119.15 (2018): 2239-2249.

[16] Lakshmanaprabu SK, K. Shankar, Ashish Khanna, Deepak Gupta, Joel J. P. C. Rodrigues, Plácido R. Pinheiro, Victor Hugo C. de Albuquerque. Effective Features to Classify Big Data using Social Internet of Things. IEEE Access, 6 (2018): 24196-24204. https://doi.org/10.1109/ACCESS.2018.2830651.

[17] Andino Maseleno, Alicia Y.C. Tang, Moamin A. Mahmoud, Marini Othman, Suntiaji Yudo Negoro, Soukaina Boukri, K. Shankar, Satria Abadi, Muhamad Muslihudin. "The Application of Decision Support System by Using Fuzzy Saw Method in Determining the Feasibility of Electrical Installations in Customer's House". International Journal of Pure and Applied Mathematics, 119. 16 (2018): 4277-4286.

[18] Mohamed Elhoseny, K. Shankar, S. K. Lakshmanaprabu, Andino Maseleno, N. Arunkumar. Hybrid optimization with cryptography encryption for medical image security in Internet of Things. Neural Computing and Applications. 2018. https://doi.org/10.1007/s00521-018-3801-x.

[19] K. Shankar, Mohamed Elhoseny, E. Dhiravida chelvi, SK. Lakshmanaprabu, Wanqing Wu. An Efficient Optimal Key Based Chaos Function for Medical Image Security. IEEE Access. 2018. https://doi.org/10.1109/ACCESS.2018.2874026.

[20] T. Avudaiappan, R. Balasubramanian, S. Sundara Pandiyan, M. Saravanan, S. K. Lakshmanaprabu, K. Shankar, "Medical Image Security Using Dual Encryption with Oppositional Based Optimization Algorithm, Journal of Medical Systems", 42.11 (2018) 1-11. https://doi.org/10.1007/s10916-018-1053-z.

[21] K.Shankar and P.Eswaran. "RGB Based Multiple Share Creation in Visual Cryptography with Aid of Elliptic Curve Cryptography", China Communications, 14. 2 (2017): 118-130. https://doi.org/10.1109/CC.2017.7868160.

[22] LakshmanaprabuS.K, Sachi Nandan Mohanty, K. Shankar, Arunkumar N, GustavoRamireze. "Optimal deep learning model for classification of lung cancer on CT images", Future Generation Computer Systems. 2018. https://doi.org/10.1016/j.future.2018.10.009.

[23] K. Shankar, Lakshmanaprabu S.K, Deepak Gupta, Andino Maseleno, Victor Hugo C. de Albuquerque. Optimal Features Based Multi Kernel SVM Approach for Thyroid Disease Classification. The Journal of Supercomputing, 2018. https://doi.org/10.1007/s11227-018-2469-4 https://doi.org/10.1007/s11227-018-2469-4.