Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.1, 2019, pp. 42-47

A Comparative Study on the Feature Selection Techniques for Intrusion Detection System

D. Selvamani[1] and V. Selvi[2]
[1]Research Scholar, [2]Assistant Professor
[1&2]Department of Computer Science, Mother Teresa Women's University, Tamil Nadu, India
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. Feature selection is employed to diminish the number of features in various applications where data has more than hundreds of attributes. Essential or relevant attribute recognition has converted a vital job to utilize data mining algorithms efficiently in today world situations. This article describes the comparative study on the Information Gain, Gain Ratio, Symmetrical Uncertainty, Chi-Square analysis feature selection techniques with different Classification methods like Artificial Neural Network, Naïve Bayes and Support Vector Machine. In this article, different performance metrics has utilized to choose the appropriate Feature Selection method for better data classification in IDS.**
Keywords: **Intrusion Detection, Feature Selection, Information Gain, Gain Ratio, Symmetrical Uncertainty, Chi-Square, Classification, ANN, Naïve Bayes and Support Vector Machine**

I. INTRODUCTION

Internet exhibits a demanding part in this moderate world. It is practiced in the shopping, education, social networking, business, 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. Principally, IDS design [2] 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 a spell prepared on the network with the numerous innovative fears in a network.

II. 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 Ibne Reaz [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.

III. FEATURE SELECTION TECHNIQUES

FS methods have listed as filter and wrapper [8]. Upon this essential concept, many FS approaches have incorporated in machine learning (ML) paradigm. Wrapper method is utilized to decide the features detected on the precision evaluation, and filter method is employed to select the features not based on the precision evaluation; instead, it uses the data features with the relevancy or correlation. Filter-based systems are not reliant on classifiers and usually quicker and extra scalable than wrapper-based methods. Moreover, they have weak computational complexity too. Recently, amounts of hybrid methods are also being introduced to achieve appropriate stability in the feature selection standards by combing both filter and wrapper method [9].

A. Information Gain Feature Selection Technique

Entropy is frequently used in the information theory measure, which exemplifies the transparency of a random collection of samples [10]. It is in the establishment of Gain Ratio, Information Gain and Similarity Uncertainty (SU) [11]. 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 [10] 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.

B. Gain Ratio Feature Selection Technique

The Gain Ratio [10] is the non-symmetrical measure that is introduced to compensate for the bias of the Information Gain (IG). GR is given by

$$GR = \frac{Information\ Gain\ (IG)}{H(A)}$$

Information Gain (IG) is a symmetrical measure.

$$IG = H(B) - H(B|A) = H(A) - H(A|B)$$

The information gained about B after observing A is equal to the information gained about A after observing B. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative. As in the equation (3.5) presents, when the variable B has to be predicted, then normalize the IG by dividing by the entropy of *A*, and vice versa. Due to this normalization, the GR values always fall in the range [0, 1]. A value of GR = 1 indicates that the knowledge of *A* completely predicts *B*, and GR = 0 means that there is no relation between *B* and *A*. In opposition to IG, the GR favors variables with fewer values.

C. Chi-Square Analysis

Feature Selection via chi-square χ^2test [12] is another, very commonly used method. Chi-squared attribute evaluation evaluates the worth of a feature by computing the value of the chi-squared statistic for the class. The initial hypothesis H_0 is the assumption that the two features are unrelated, and the chi-squared formula tests it:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\frac{O_{ij} - E_{ij}}{E_{ij}}\right)^2 \quad (3.4)$$

Where O_{ij} is the observed frequency, and E_{ij} is the expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of χ^2, the greater the evidence against the hypothesis H_0.

D. Symmetrical Uncertainty Feature Selection Method

The symmetrical uncertainty (SU) [13] among the features and target concept are applied to achieve the best features for classification. The features with greater SU values have the more significant weight. SU estimates the association among R, S features based on the information theory [14]. It will calculate as follows

$$SU\ (R,)\ S = 2\ \frac{I\ (R,S)}{H(SR) + H\ (S)}$$

Estimating I(R, S) as the MI [14] among R, S. H(..) as an entropy function for R, S features. The SU means the [0,1] (normalized range value) as the improvement factor value is 2. The value of SU for one feature is 1, and then it is predictable. The value of SU is 0, then R and S features do not have the relationship.

IV. DESCRIPTION OF THE DATASET

Aware of the deficiency of appropriate audit data sets for intrusion detection, KDDCUP [15] 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 I DESCRIPTION OF THE KDD CUP 99 DATASET

S. 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

V. RESULTS AND DISCUSSION

Following table II provides the outcome attained by the proposed Precocious Feature Selection method and current filter-based feature selection techniques like Gain Ratio, Symmetrical Uncertainty, Chi-Square analysis and Information Gain. From table II, Gain Ratio filters 32 features, Symmetrical Uncertainty filters 31 features, Information Gain screens only 27 features, and the Chi-Square gives only 32 features. To assess the competence of the filter feature selection 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 II NUMBER OF FEATURES OBTAINED BY GAIN RATIO, SYMMETRICAL UNCERTAINTY, CHI-SQUARE ANALYSIS AND I NFORMATION GAIN FEATURE SELECTION TECHNIQUES

S. No.	Feature Selection Techniques			
	Gain Ratio	**Symmetrical Uncertainty**	**Chi-Square Analysis**	**Information Gain**
1	is_guest_32	src_bytes	dst_host_srv_count	Src_bytes
2	logged_in	dst_bytes	Dst_bytes	dst_host_same_srv_rate
3	Dst_bytes	dst_host_srv_count	dst_host_same_srv_rate	Dst_bytes
4	dst_host_srv_serror_rate	dst_host_same_srv_rate	srv_Count	dst_host_rerror_rate
5	Src_bytes	dst_host_rerror_rate	count	Service
6	dst_host_serror_rate	Service	dst_host_count	dst_host_diff_srv_rate
7	dst_host_same_srv_rate	dst_host_srv_rerror_rate	duration	dst_host_srv_rerror_rate
8	dst_host_srv_count	dst_host_serror_rate	dst_host_rerror_rate	srv_Count
9	dst_host_srv_rerror_rate	logged_in	dst_host_same_src_port_rate	count
10	dst_host_rerror_rate	dst_host_srv_serror_rate	Service	dst_host_serror_rate
11	Service	dst_host_diff_srv_rate	dst_host_srv_rerror_rate	dst_host_srv_serror_rate
12	dst_host_count	srv_Count	dst_host_diff_srv_rate	dst_host_same_src_port_rate
13	hot	dst_host_count	dst_host_srv_diff_host_rate	logged_in
14	dst_host_diff_srv_rate	dst_host_same_src_port_rate	srv_serror_rate	dst_host_count
15	srv_rerror_rate	count	srv_diff_host_rate	rerror_rate
16	dst_host_same_src_port_rate	flag	diff_srv_rate	srv_rerror_rate
17	Flag	srv_rerror_rate	serror_rate	same_srv_rate
18	srv_count	rerror_rate	dst_host_serror_rate	diff_srv_rate
19	rerror_count	is_guest_32	same_srv_rate	dst_host_srv_diff_host_rate
20	dst_host_srv_diff_host_rate	dst_host_srv_diff_host_rate	rerror_rate	serror_rate
21	srv_serror_rate	same_srv_rate	hot	srv_serror_rate
22	count	diff_srv_rate	Protocol_type	is_guest_32
23	serror_rate	hot	srv_rerror_rate	hot
24	diff_srv_rate	srv_serror_rate	Flag	Protocol_type
25	Protocol_type	hot	dst_host_srv_serror_rate	srv_diff_host_rate
26	same_srv_rate	srv_serror_rate	Wrong_fragment	num_compromised
27	srv_diff_host_rate	serror_rate	num_failed_32s	num_failed_32s
28	num_compromised	Protocol_type	num_compromised	
29	num_failed_32s	srv_diff_host_rate	is_guest_32	
30	Wrong_fragment	num_compromised	logged_in	
31	land	num_failed_32s	land	
32	urgent		Urgent	

Following table III gives the performance analysis of the original dataset, Gain Ratio, Symmetrical Uncertainty, Chi-Square analysis and Information Gain Feature Selection method by using Artificial Neural Network as the classifier. From table III it is strong that the Information Gain technique performs effectively and contributes the maximized accuracy, Kappa Statistics, TPR, FPR, Precision, Recall, F-Measure, and ROC Area.

TABLE III PERFORMANCE ANALYSIS OF THE ORIGINAL DATASET, INFORMATION GAIN, GAIN RATIO AND CHI-SQUARE ANALYSIS FEATURE SELECTION USING ARTIFICIAL NEURAL NETWORK CLASSIFICATION METHOD

Evaluation Metrics	Original Dataset	Feature Selection Techniques			
		Gain Ratio	Symmetrical Uncertainty	Chi-Square analysis	Information Gain
Accuracy	69.3333 %	70.4136 %	71.4136%	74.5921 %	93.7677 %
Relative Absolute Error (RRAE)	45.3867 %	45.2186%	45.1124%	44.4213%	43.384 %
Root Relative Squared Error (RRSE)	88.8892 %	86.7215%	87.8106 %	86.2340%	77.864 %
Kappa Statistics	0.5539	0.5612	0.5616	0.5891	0.6242
True Positive Rate	0.682	0.701	0.732	0.831	0.938
False Positive Rate	0.16	0.15	0.16	0.191	0.283
Precision	0.642	0.59	0.601	0.713	0.934
Recall	0.682	0.701	0.732	0.831	0.938
F-Measure	0.612	0.632	0.631	0.781	0.942
ROC Area	0.836	0.841	0.839	0.840	0.868

TABLE IV PERFORMANCE ANALYSIS OF THE ORIGINAL DATASET, INFORMATION GAIN, GAIN RATIO AND CHI-SQUARE ANALYSIS FEATURE SELECTION USING NAÏVE BAYES CLASSIFICATION METHOD

Evaluation Metrics	Original Dataset	Feature Selection Techniques			
		Gain Ratio	Symmetrical Uncertainty	Chi-Square analysis	Information Gain
Accuracy	60.9091 %	69.2415%	65.4545 %	66.3838 %	71.2415%
Root Relative Absolute Error (RRAE)	91.0323 %	86.9231 %	85.4265 %	87.3826 %	76.9231 %
Root Relative Squared Error (RRSE)	98.9294 %	97. 0277 %	96.4227 %	95.3548 %	93.6921%
Kappa Statistics	0.1508	0.3528	0.3842	0.4176	0.4489
True Positive Rate	0.59	0.761	0.655	0.609	0.773
False Positive Rate	0.239	0.41	0.375	0.457	0.201
Precision	0.575	0.701	0.668	0.457	0.771
Recall	0.59	0.761	0.655	0.609	0.773
F-Measure	0.582	0.721	0.659	0.61	0.766
ROC Area	0.746	0.715	0.713	0.656	0.764

TABLE V PERFORMANCE ANALYSIS OF THE ORIGINAL DATASET, INFORMATION GAIN, GAIN RATIO AND CHI-SQUARE ANALYSIS FEATURE SELECTION USING SUPPORT VECTOR MACHINE CLASSIFICATION METHOD

Evaluation Metrics	Original Dataset	Feature Selection Techniques			
		Gain Ratio	Symmetrical Uncertainty	Chi-Square analysis	Information Gain
Accuracy	66.4312 %	67.6346 %	68.1240%	69.4123%	75.3333 %
Relative Absolute Error (RRAE)	92.8816 %	99.818 %	98.212%	98.921%	97.2935 %
Root Relative Squared Error (RRSE)	93.562 %	99.999 %	98.781%	97.256%	88.2774 %
Kappa Statistics	0.3553	0.3846	0.3689	0.3741	0.4118
True Positive Rate	0.642	0.672	0.631	0.635	0.742
False Positive Rate	0.22	0.34	0.39	0.376	0.26
Precision	0.494	0.651	0.635	0.645	0.647
Recall	0.642	0.672	0.631	0.635	0.742
F-Measure	0.545	0.653	0.632	0.634	0.691
ROC Area	0.8	0.622	0.603	0.645	0.822

Following table IV gives the performance analysis of the original dataset, Gain Ratio, Symmetrical Uncertainty, Chi-Square analysis and Information Gain Feature Selection method by using Naïve Bayes as the classifier. From table

IV it is strong that the Information Gain technique performs effectively and contributes the maximized accuracy, Kappa Statistics, TPR, FPR, Precision, Recall, F-Measure, and ROC Area.

Following table V gives the performance analysis of the original dataset, Gain Ratio, Symmetrical Uncertainty, Chi-Square analysis and Information Gain Feature Selection method by using Support Vector Machine as the classifier. From table V it is strong that the Information Gain technique performs effectively and contributes the maximized accuracy, Kappa Statistics, TPR, FPR, Precision, Recall, F-Measure, and ROC Area.

VI. 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 article, various filtered feature selection approach has used to heighten the accuracy of the classification in the IDS. The Feature selection methods have 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 Information Gain feature selection 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.

[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.

[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.

[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.

[5] Aburomman, Abdulla Amin and Mamun Bin Ibne Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," *Applied Soft Computing,* Vol. 38, pp. 360-372, 2016.

[6] Thaseen, Ikram Sumaiya and Cherukuri Aswani 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.

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

[8] Pereira, Rafael B., *et al.*, "Categorizing feature selection methods for multi-label classification", *Artificial Intelligence Review,* Vol. 49, No. 1, pp. 57-78, 2018.

[9] Sheikhpour, Razieh, *et al.*, "A survey on semi-supervised feature selection methods", *Pattern Recognition,* Vol. 64, pp. 141-158, 2017.

[10] Jadhav, Swati, Hongmei He, and Karl Jenkins, "Information gain directed genetic algorithm wrapper feature selection for credit rating", *Applied Soft Computing,* Vol. 69, pp. 541-553, 2018.

[11] Venkataraman, Sivakumar, and Rajalakshmi Selvaraj, "Optimal and Novel Hybrid Feature Selection Framework for Effective Data Classification", *Advances in Systems, Control and Automation*, *Springer*, Singapore, pp. 499-514, 2018.

[12] Thaseen, Ikram Sumaiya, and Cherukuri Aswani 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.

[13] Moayedikia, Alireza, *et al.*, "Feature selection for high dimensional imbalanced class data using harmony search", *Engineering Applications of Artificial Intelligence*, Vol. 57, pp. 38-49, 2017.

[14] Wang, Yintong, et al., "An efficient semi-supervised representatives feature selection algorithm based on information theory", *Pattern Recognition,* Vol. 61, pp. 511-523, 2017.

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