RANDOM FOREST CLASSIFICATION OF NSL-KDD DATASET USING HYBRID FEATURE SELECTION MODEL

[1]Priyanka Sharma, [2]Rajni Ranjan Singh Makwana
[1]Student, Department of CSE & IT, Madhav Institute of Technology & Science, Gwalior, India
[2]Assistant Professor, Department of CSE & IT, Madhav Institute of Technology & Science, Gwalior, India

Abstract: IDS is a security software that detects a network for malicious activity or policy violations. An IDS work by monitoring system for known attack patterns and utilized to identify the types of attack. In this work a hybrid feature selection model has been introduced to perform random forest classification of NSL-KDD dataset. Here 20 unique combinations of wrapper and filter feature selection methods have been created to select relevant attributes. To analyze the effectiveness and usefulness of proposed way an observation has been accomplished using WEKA machine learning tool. It is observed that filter method based gain ratio feature selection method gives better result that is 99.46% (in 10 fold cross validation) and 99.99% (in use training set).

KEYWORDS: *Intrusion Detection System, NSL-KDD dataset, Random Forest Classification Algorithm, Feature Selection Techniques.*

1. INTRODUCTION

An intrusion detection system detects network traffic and detector for distrustful activity and alert the system. It is originally built for vulnerebility in the system and it alerts when any type of activity discovered. There are basically two methods for IDS: HIDS OR NIDS.HIDS takes the snapshots of system's file set and matches it to a last snapshot. HIDS runs on separate devices on the network such as firewalls and antivirus software etc. NIDS can monitor incoming, outgoing and local traffic. NIDS installed only at specific point such as servers. It does analysis for network traffic on entire subnet. IDS has mainly two approaches: signature detection and anomaly detection.Signature detection is a pattern or mark is compared with past events to find current threats .Anomaly detection protects against unknown threats. In the event if any movement is observed to be strange from the baseline, the alarm is activated by the IDS associated with an interruption. The baseline depicts the normal behaviour of the traffic.

1.1 LITERATURE REVIEW

Abebe tesfahun et. al. [3] proposed a model in which NSL-KDD full training dataset has been utilized and 10 fold cross validation used for the testing purpose. Smote with minority classes (r2l and u2r) have been selected for solving the problem of imbalanced training data. Information gain feature selection has been applied on smote generated dataset for feature reduction. The result shows that the detection rate has increased for the minority class (u2r) and this approach also reduces the time.

Prashant kushwaha et. al. [4] proposed an algorithm for recognition of an attack from the normal traffic. In this paper various types of feature selection algorithm used for finding best attributes from 10%kdd-cup99 dataset. on the basis of mutual information 30 attributes selected and various classificification algorithms applied and found that SVM gave the best result among all other classifiers.SVM improves accuracy to(99.91%).

Himadri chauhan et. al. [5] presented the application of data mining to IDS. In this paper 20% NSL-KDD dataset were used and top 10 classification algorithm were applied by utilizing 10 fold cross validation to test. The result was shown that random forest classifier has got first position with 99.75% classification accuracy.

Nutan farah haq et. al. [6] proposed a framework with a hybrid feature selection using 20% NSL-KDD training dataset. Three different classifiers utilized for classification with different wrapper based search methods. Final features chose by different wrapper search techniques for three different classifiers. The result depicts that naïve bayes classifiers works best with BFS technique ,Bayesian network classifier works best with genetic search and j48 classifier works better with genetic search technique. Final feature set used an ensamble approach using majority vote concept which used (Naïve bayes, Bayesian network and j49 as the base classifier) and this approach gives high accuracy and best performer than the best combinational setup.

Tanya garg et. al. [7] presented a paper in which NSL-KDD dataset using 10 classification algorithm with five cross validation. So observed that rotation forest classification algorithm have highest accuracy among all classifiers. The combination of feature selection techniques were used and observed that symmetric and gain with IBK classification perform better with high accuracy.

Tanya garg et. al. [12] presented the comparative performance of NSL-KDD dataset compatible classification algorithms. According to their performance Garret's ranking have been applied to rank these classification algorithms and It is observed that rotation forest classifier gives better result.

Ahmad riza'ain yusof et. al. [14] presented a paper to proposed combining of two feature selection techniques using DCF and CSE to identify most relevant features using NSL-KDD dataset. These two techniques consider with the traditional feature selection techniques such as IG, GR, chi-squared and CFS. The Proposed method gives high accuracy in all aspect compared to other methods.

Sumaiya thaseen et. al. [15] presented a paper to classify network events in intrusion detection system by evaluated eight tree based classifications algorithms. NSL-KDD dataset used for analysis and observed that random tree classifier gives high accuracy and reduced false alarm rate.

2. DATASET DESCRIPTION

NSL-KDD (Network security lab-knowledge discovery in databases) is one of the first research datasets for network anomaly detection. The data set name was changed to NSL-KDD by detecting and correcting false errors in the dataset. It consists of 41 attributes values and each of these instances are classified either normal or the attack type [1]. NSL-KDD dataset contains 25,192 records.

The NSL-KDD dataset includes following benefits.

1. It does not include redundancy of data or duplicate records in the train set, so the classifier will not be biased towards more repeated records.

2. The number of records in the train and test sets is sensible ,which makes it reasonable to run the experiments on the total set without the need to randomly choose a little portion [2].

TABLE-I

NSL-KDD 20% Training Dataset

NSL- KDD 20% Training Dataset **Attribute : 42**	
Label	Count
Normal	13,499
Anomaly	11,743
Total	25,192

2.1 Random forest classification:

It is a decision tree based algorithm. Compared to a single decision tree algorithm random forest runs efficiently on large data sets with a better accuracy [3] . Random forest is the best classification algorithm compared to others. This algorithm gives high accuracy.

Shashikant upadhyay et. al. [16] proposed an attack specific classification of KDD cup dataset. Mainly five classifiers were utilized to determine better classifiers for each in different attacks based on precision, recall, F-measure and ROC curve area performance criteria's. It was found that naive bayes performed better in terms of false positive rate. And it was also found that random forest is the best classifier among all used methods.

Jasy elsa Varghese et. al. [13] focused on comparison of accuracy of individual classifiers with two feature selection methods using NSL-KDD dataset, where random forest with PCA gives high accuracy rate of 99.52%.

3. PROPOSED WORK

In this paper a Model has been proposed to perform feature selection of NSL-KDD dataset using random forest classification. Random forest classification has been applied on reduced features set after hybrid feature selection using 10 fold cross validation and use training set. Analysis is accomplished on the basis of classification accuracy. This proposed work has mainly two phases which is depicted in fig. 1

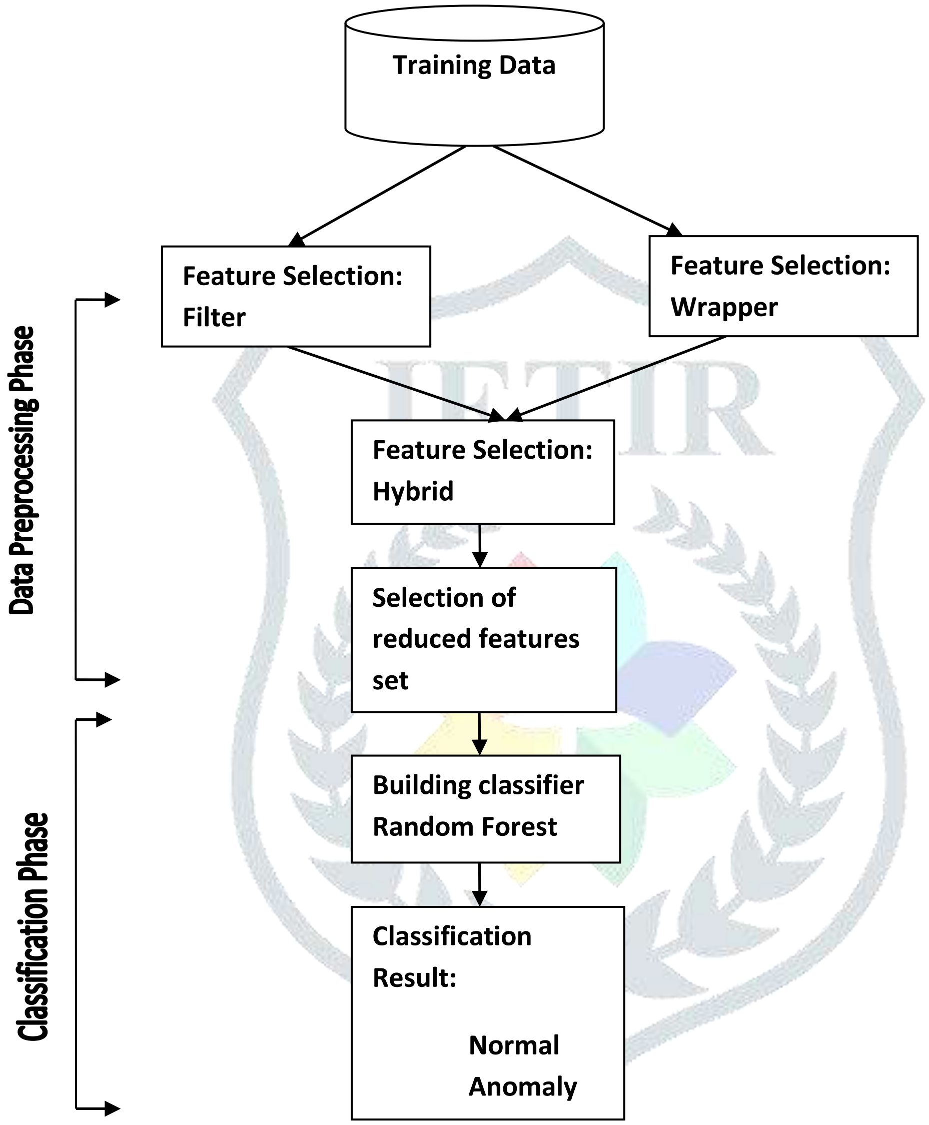
Training Data
Feature Selection: Filter
Feature Selection: Wrapper
Feature Selection: Hybrid
Selection of reduced features set
Building classifier Random Forest
Classification Result:
Normal
Anomaly
Data Preprocessing Phase
Classification Phase

Fig. 1 Hybrid feature selection model

Phase 1. Data pre-processing phase- In this phase NSL-kDD dataset has been utilized. Reduced features set have been selected by applying the approach of hybrid feature selection.

Phase 2: Training phase- In this phase random forest classification applied on reduced features set after hybrid feature selection of NSL-KDD dataset.

3.1 Feature Selection

Feature selection is a method for removing irrelevant and redundant features and choosing the most optimal subset of features that create a good characterization of pattern belonging to individual classes [9]. It is extremely important and frequently used technique data pre-processing for data mining.

There are basically two techniques for feature selection: Wrapper and Filter. The wrapper techniques uses a subset evaluator. This will create all feasible subsets from feature vector. Then it will use a classification algorithm to encourage classifiers from the features in each subset.

The filter method uses an attribute evaluator and a ranker to rank all the features in dataset. The no. of features that you want to choose from your feature vector can always be defined.

In this paper, hybrid feature selection applied as a pre-processing step. It improves accuracy, precision, computational time etc. In our research three wrapper based techniques are used:

3.1.1. Cfs Subset Evaluator-It stands for cor-relation based feature selection .Cfs is used to select suitable feature while eliminating redundant ones. Cfs uses a subset evaluator instead of different attributes and is able to finding usefull attributes by observe features and relevancy between features and class label [10] .

3.1.2. Consistency Subset Evaluator-This technique is used to find optimal and evaluate a subset of suitable features that are consistent with each other.CSE identified by representing a combination of feature value with a class given pattern label. And the given pattern of features should produce the same class [8] .

3.1.3. Filtered subset evaluator-This technique measure the usefulness of a subset of feature. This method can provide the best subset of features.

Three filter based techniques are used in our research paper.

3.1.4 OneR-It is a rule based algorithm as it generates rules and according of those rules it chooses features and rank them accordingly [7] .

3.1.5. Gain Ratio-Gain ratio is a modification in information gain to decrease its biasness. It selects the number of branches when taking an appropriate attribute. It is based on the information given by intrinsic attribute. It selects an attribute on the basis of intrinsic information. Intrinsic information is the information to determine which branch belongs to which instance [7] .

3.1.6. Information gain- Information gain is based on the idea of entropy. The major drawback of using information gain is it selects attribute with large numbers of separate values over attributes with fewer values [7] .

3.2 Performance Metrices

To evaluate the performance of proposed model following performance criteria has been utilized.

TABLE-II CONFUSION MATRIX

		Positive	**Negative**
Actual	Positive	A: True Positive	B: False Negative
	Negative	C: False Positive	D: True Negative

3.2.1. True positive rate/Recall- It is the part of cases when results are positive and that were precisely classified as positive, as deliberated by utilizing the following equation [11].

$$\mathbf{Recall=A/A+B}$$

3.2.2. False Positive Rate- It is the part of cases whose results should be negative but that were imprecisely classified positive, as calculated by utilizing [11]

$$\mathbf{FP\ Rate=C/C+D}$$

3.2.3. Precision- It is the section of the predicted positive cases which were correct and as delibrated utilizing by following equation [11].

$$\mathbf{Precision=A/A+C}$$

3.2.4. F-Measure- The F-Measure estimates some mean of all the information retrieval precision and recall metrices [11].

F = 2. Precision * recall/ Precision + recall

3.2.5. ROC Curve-This curve describes that how excellent good and worthless experiments are plotted on the same graph [11].

3.2.6. Kappa Statics-Kappa statics is used to measure the accordance in between predicted and observed categorization of a dataset, while correcting for an accordance that happens by coincidence. If the results of kappa is 1 then it specifies precisely accordance where as if the result of kappa is 0 then is specifies accordance equals to chance [11].

3.2.7. Classification – It is conditional on the number of samples exactly classified. Here t is the number of sample cases correctly classified and n is the total number of sample cases [11].

Classification % =100*t/n

4. EXPERIMENTATION AND RESULT

In this paper hybrid feature selection approach has been applied to NSL-KDD dataset using random forest classification. To measure the performance of effectiveness of proposed work an experiment has been carried out using well known weka machine learning tool version 3.6. Here hybrid feature selection approach has been using three wrapper based Cfs subset evaluator, Consistency subset evaluator and filtered subset evaluator techniques are used and three filter based techniques OneR, Gain Ratio and Information gain methods. The classification has been carried out using random forest classification.

Table III shows the list of reduced features set after feature selection and hybrid feature selection . Table IV show 10 Fold cross validation used for testing purpose. In table V shows use training set used for testing purpose.

TABLE –III List of reduced features set using various hybrid feature techniques

Feature Selection Techniques	**Selected Features**
CfS + Best First	4,5,6,12,26,29,30
Consistency+ Best First	1,3,4,5,14,23,32,34,35,37
Filtered+ Best First	4,5,6,12,26,29,30
OneR+ Ranker	5,3,6,4,29,30,34,33,35,12,23,25,38,39,26,32,36,37,24, 31,41,40,27,28,2,8,10,13,1,14,20,22,18,19,21,9,15,16, 7,17,11
Gain Ratio+ Ranker	12,26,4,25,39,6,30,38,5,29,3,37,34,33,8,35,23,31,41,32, 28.27,36,16,15,2,10,13,19,1,40,18,17,24,14,22,7,11,20, 9,21
Info gain+ Ranker	5,3,6,4,30,29,33,34,35,38,12,39,25,23,26,37,32,36,31, 24,41,2,27,40,28,1,10,8,13,16,19,22,17,15,14,18,7,11,9, 20,21
CfS + Consistency+ Filtered	4,5,37
OneR+ Gain Ratio+ Info Gain	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21, 22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39 40,41
Cfs+ Consistency + OneR	4,5,37
Consistency+ Filtered+ OneR	4,5,37
Filtered +Cfs+ OneR	4,5,6,12,26,29,30,37
OneR+ Gain Ratio+ Cfs	4,5,6,12,26,29,30
Gain Ratio+ Info Gain+ Cfs	4,5,6,12,26,29,30
Info Gain+ OneR+ Cfs	4,5,6,12,26,29,30
OneR+ Gain Ratio+ Consistency	1,3,4,5,14,23,32,34,35,37
Gain Ratio+ Info Gain+ Consistency	1,3,4,5,14,23,32,34,35,37
Info Gain+ OneR+ Consistency	1,3,4,5,14,23,32,34,35,37
OneR+ Gain Ratio+ Filtered	4,5,6,12,26,29,30
Gain Ratio+ Info Gain+ Filtered	4,5,6,12,26,29,30
Info Gain+ OneR+ Filtered	4,5,6,12,26,29,30

TABLE-IV Experiment Using 10 Fold Cross Validation

Name of Techniques	Classifier	Training Time	Accuracy	FPR	ROC	Recall	Preci-sion	Error	Kappa
CFS+ Best first	RANDOM FOREST	21.77 sec	97.6421	0.007	0.998	0.976	0.974	0.0067	0.9579
Consistency+ Best First	RANDOM FOREST	34.86 sec	92.3865	0.026	0.984	0.924	0.918	0.0178	0.8627
Filtered+ Best First	RANDOM FOREST	23.3 sec	97.6421	0.007	0.998	0.976	0.974	0.0067	0.9579
OneR+ Ranker	RANDOM FOREST	41.14 sec	99.4601	0.003	1	0.995	0.994	0.0024	0.9904
Gain Ratio+ Ranker	RANDOM FOREST	39.71 sec	99.4601	0.003	1	0.995	0.994	0.0024	0.9904
Info Gain+ Ranker	RANDOM FOREST	45.65 sec	99.4601	0.003	1	0.995	0.994	0.0024	0.9904
CFS+ Consistency+ Filtered	RANDOM FOREST	34.31 sec	88.4527	0.059	0.956	0.885	0.872	0.0323	0.7844
OneR+ Gain Ratio +Info Gain	RANDOM FOREST	41.52 sec	99.4601	0.003	1	O.995	0.994	0.0024	0.9904
CFS+ Consistency+ OneR	RANDOM FOREST	17.22 sec	88.4527	0.059	0.956	0.885	0.872	0.0323	0.7844
Consistency+ Filtered+ OneR	RANDOM FOREST	17.96 sec	88.4527	0.003	1	0.995	0.994	0.0024	0.9904
Filtered+ CFS+ OneR	RANDOM FOREST	25.1 sec	97.6421	0.007	0.998	0.976	0.974	0.0067	0.9579
OneR+ Gain Ratio +CFS	RANDOM FOREST	22.23 sec	97.6421	0.007	0.998	0.976	0.974	0.0067	0.9579
Gain ratio+ info gain +CFS	RANDOM FOREST	21.47 sec	97.6421	0.007	0.998	0.976	0.974	0.0067	0.9579
Info Gain+ OneR +CFS	RANDOM FOREST	22.31 sec	97.6421	0.007	0.998	0.976	0.974	0.0067	0.9579
OneR+ Gain Ratio+ Consistency	RANDOM FOREST	34.19 sec	92.3865	0.026	0.984	0.924	0.918	0.0178	0.8627
Gain Ratio+ Info Gain+ Consistency	RANDOM FOREST	14.33 sec	92.3865	0.026	0.984	0.924	0.918	0.0178	0.8627
Info Gain+ OneR+ Consistency	RANDOM FOREST	33.46 sec	92.3865	0.026	0.984	0.924	0.918	0.0178	0.8627
OneR+ Gain Ratio +Filtered	RANDOM FOREST	21.44 sec	97.6421	0.007	0.998	0.976	0.974	0.0067	0.9579
Gain Ratio +Info Gain +Filtered	RANDOM FOREST	25.44 sec	97.6421	0.007	0.998	0.976	0.974	0.0067	0.9579
Info Gain+ OneR+ Filtered	RANDOM FOREST	22.17 sec	97.6421	0.007	0.998	0.976	0.974	0.0067	0.9579

TABLE-V Experiment Using Use Training Set

Name of Techniques	Classifier	Training Time	Accuracy	FPR	ROC	Recall	Preci-sion	Error	Kappa
CFS+ Best First	RANDOM FOREST	20.33 sec	98.1224	0.002	0.999	0.981	0.982	0.0057	0.9665
Consistency + Best First	RANDOM FOREST	29.29 sec	99.3609	0.003	1	0.994	0.994	0.0071	0.9886
Filtered+ Best First	RANDOM FOREST	27.71 sec	98.1224	0.002	0.999	0.981	0.982	0.0057	0.9665
OneR+ Ranker	RANDOM FOREST	36.44 sec	99.996	0	1	1	1	0.008	0.9999
Gain Ratio+ Ranker	RANDOM FOREST	32.49 sec	99.996	0	1	1	1	0.008	0.9999
Info gain+ Ranker	RANDOM FOREST	38.31 sec	99.996	0	1	1	1	0.008	0.9999
CFS+ Consistency+ Filtered	RANDOM FOREST	15.62 sec	89.0084	0.054	0.965	0.89	0.897	0.0314	0.795

OneR+ Gain Ratio+ Info Gain	RANDOM FOREST	36.33 sec	99.996	0	1	1	1	0.008	0.9999
CFS +Consistency+ OneR	RANDOM FOREST	18.6 sec	89.0084	0.054	0.965	0.89	0.897	0.0314	0.9999
Consistency+ Filtered+ OneR	RANDOM FOREST	17.27 sec	89.0084	0.054	0.965	0.89	0.897	0.0314	0.9999
Filtered +CFS+ OneR	RANDOM FOREST	16.38 sec	98.1224	0.002	0.999	0.981	0.982	0.0057	0.9665
OneR +Gain Ratio +CFS	RANDOM FOREST	20.7 sec	98.1224	0.002	0.999	0.981	0.982	0.0057	0.9665
Gain Ratio+ Info Gain +CFS	RANDOM FOREST	20.8 sec	98.1224	0.002	0.999	0.981	0.982	0.0057	0.9665
Info Gain+ OneR +CFS	RANDOM FOREST	20.58 sec	98.1224	0.002	0.999	0.981	0.982	0.0057	0.9665
OneR +Gain Ratio +Consistency	RANDOM FOREST	28.61 sec	99.3609	0.003	1	0.994	0.994	0.0071	0.9886
Gain Ratio+ Info Gain+ Consistency	RANDOM FOREST	38.61 sec	99.3609	0.003	1	0.994	0.994	0.0071	0.9886
Info Gain+ OneR+ Consistency	RANDOM FOREST	28.61 sec	99.3609	0.003	1	0.994	0.994	0.0071	0.9886
OneR+ Gain Ratio+ Filtered	RANDOM FOREST	25.92 sec	98.1224	0.002	0.999	0.981	0.982	0.0057	0.9665

TABLE- VI Experiment Summary of 10 Fold Cross Validation

Performance Metrices	Name of Techniques	Low	Name of Techniques	High
ACCURACY	CFS+ Consistency+ Filtered	88.45	OneR+ Ranker	99.46
	CFS+ Consistency+ OneR		GR+ Ranker	
	Consistency+ Filtered+ OneR		IG+ Ranker	
			OneR+ GR+ IG	
TRAINING TIME	CFS+ Consistency+ Filtered	34.31	OneR+ Ranker	41.74
	CFS+ Consistency+ OneR	17.22	GR+ Ranker	39.71
	Consistency+ Filtered+ OneR	17. 96	IG+ Ranker	45.65
			OneR+ GR+ IG	41.52
FPR	OneR+ Ranker	0.003	CFS+ Consistency+ Filtered	0.059
	GR+ Ranker		CFS+ Consistency+ OneR	
	IG+ Ranker		Consistency+ Filtered+ OneR	
	OneR+ GR+ IG			
ROC	CFS+ Consistency+ Filtered	0.95	OneR+ Ranker	1
	CFS+ Consistency+ OneR		GR+ Ranker	
	Consistency+ Filtered+ OneR		IG+ Ranker	
			OneR+ GR+ IG	
RECALL	CFS+ Consistency+ Filtered	0.88	OneR+ Ranker	0.99
	CFS+ Consistency+ OneR		GR+ Ranker	
	Consistency+ Filtered+ OneR		IG+ Ranker	

			OneR+ GR+ IG	
PRECISION	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	0.87	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	0.99
ERROR	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	0.02	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	0.032
KAPPA	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	0.78	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	0.99

Experiment summary of 10 fold cross validation is mentioned in table VI. From table VI it is observed that filter based feature selection technique gain ratio with ranker search method gives best performance in terms of classification accuracy, precision, recall, Training time and showing low FPR and error. And it is also observed that wrapper based feature selection and combining wrapper with filter method gives worst performance in all aspect and showing highest FPR and error.

TABLE- VII Experiment Summary of Use Training Set

Performance Metrices	**Name of Techniques**	**Low**	**Name of Techniques**	**High**
ACCURACY	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	89.00	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	99.99
TRAINING TIME	CFS+ Consistency+ Filtered	15.62	OneR+ Ranker	36.44
	CFS+ Consistency+ OneR	18.6	GR+ Ranker	32.49
	Consistency+ Filtered+ OneR	17.27	IG+ Ranker	38.31
			OneR+ GR+ IG	36.33
FPR	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	0	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	0.05
ROC	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	0.965	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	1
RECALL	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	0.89	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	1

PRECISION	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	0.89	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	1
ERROR	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	0.008	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	0.031
KAPPA	CFS+ Consistency+ Filtered CFS+ Consistency+ OneR Consistency+ Filtered+ OneR	0.79	OneR+ Ranker GR+ Ranker IG+ Ranker OneR+ GR+ IG	0.99

On the other side An experiment summary of use training set is mentioned in table VII. From table VII it is observed that filter based feature selection technique gain ratio with ranker search method gives better result in all aspects. And it is also observed that wrapper based hybrid feature selection and combining wrapper with filter method gives poor performance In terms of accuracy, recall, precision, roc and kappa and showing highest FPR and error.

5. CONCLUSION & FUTURE WORK

In the proposed work a hybrid feature selection technique has been applied on NSL-KDD dataset. NSL-KDD dataset has been used to perform random forest classification on reduced features set after hybrid feature selection. From both experiment summary tables it is observed that hybrid wrapper feature selection and combining wrapper with filter method gives low accuracy in all aspect. And it is also observed that filter based feature selection Gain ratio is best in all aspect. It gives high accuracy and less time consuming in both testing modes. In future, hybrid classification approach or ensemble approach may further improves the results.

References

[1] Gül, A. and Adalı, E., 2017, October. A feature selection algorithm for IDS. In *Computer Science and Engineering (UBMK), 2017 International Conference on* (pp. 816-820). IEEE.

[2] Meena, G. and Choudhary, R.R., 2017, July. A review paper on IDS classification using KDD 99 and NSL KDD dataset in WEKA. In *Computer, Communications and Electronics (Comptelix), 2017 International Conference on* (pp. 553-558). IEEE.

[3] Tesfahun, A. and Bhaskari, D.L., 2013, November. Intrusion detection using random forests classifier with SMOTE and feature reduction. In *Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), 2013 International Conference on* (pp. 127-132). IEEE.

[4] Kushwaha, P., Buckchash, H. and Raman, B., 2017, November. Anomaly based intrusion detection using filter based feature selection on KDD-CUP 99. In *Region 10 Conference, TENCON 2017-2017 IEEE* (pp. 839-844). IEEE.

[5] Chauhan, H., Kumar, V., Pundir, S. and Pilli, E.S., 2013, August. A comparative study of classification techniques for intrusion detection. In *Computational and Business Intelligence (ISCBI), 2013 International Symposium on* (pp. 40-43). IEEE.

[6] Haq, N.F., Onik, A.R. and Shah, F.M., 2015, November. An ensemble framework of anomaly detection using hybridized feature selection approach (HFSA). In *SAI Intelligent Systems Conference (IntelliSys), 2015* (pp. 989-995). IEEE.

[7] Garg, T. and Kumar, Y., 2014, December. Combinational feature selection approach for network intrusion detection system. In *Parallel, Distributed and Grid Computing (PDGC), 2014 International Conference on* (pp. 82-87). IEEE.

[8] Yusof, A.R.A., Hamdan, H., Udzir, N.I., Abdullah, M.T. and Selamat, A., 2017, November. Adaptive feature selection for denial of services (DoS) attack. In *Application, Information and Network Security (AINS), 2017 IEEE Conference on* (pp. 81-84). IEEE.

[9] Ambusaidi, M.A., He, X., Nanda, P. and Tan, Z., 2016. Building an intrusion detection system using a filter-based feature selection algorithm. *IEEE transactions on computers*, *65*(10), pp.2986-2998.

[10] Shahbaz, M.B., Wang, X., Behnad, A. and Samarabandu, J., 2016, October. On efficiency enhancement of the correlation-based feature selection for intrusion detection systems. In *Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual* (pp. 1-7). IEEE.

[11] Dubey, R. and Makwana, R.R.S., 2019. Computer-Assisted Valuation of Descriptive Answers Using Weka with Random Forest Classification. In *Proceeding of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017)* (pp. 359-366). Springer, Singapore.

[12] Garg, T. and Khurana, S.S., 2014, May. Comparison of classification techniques for intrusion detection dataset using WEKA. In *Recent Advances and Innovations in Engineering (ICRAIE), 2014* (pp. 1-5). IEEE.

[13] Varghese, J.E. and Muniyal, B., 2017, September. An investigation of classification algorithms for intrusion detection system—A quantitative approach. In *Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on* (pp. 2045-2051). IEEE.

[14] Yusof, A.R.A., Hamdan, H., Udzir, N.I., Abdullah, M.T. and Selamat, A., 2017, November. Adaptive feature selection for denial of services (DoS) attack. In *Application, Information and Network Security (AINS), 2017 IEEE Conference on* (pp. 81-84). IEEE.

[15] Thaseen, S. and Kumar, C.A., 2013, February. An analysis of supervised tree based classifiers for intrusion detection system. In *Pattern recognition, informatics and mobile engineering (prime), 2013 international conference on* (pp. 294-299). IEEE.

[16] Upadhyay, S. and Singh, R.R., 2015. Comparative Analysis based Classification of KDD'99 Intrusion Dataset. *International Journal of Computer Science and Information Security*, *13*(3), p.14.