Hybridized Combinational Feature Selection Framework for Network Intrusion Detection System (HCFSF)

Shivangi Seth
Department of CSE and IT
MITS, Gwalior, India
sethshivangi92@gmail.com

Rajni Ranjan Singh Makwana
Department of CSE and IT
MITS, Gwalior, India
rrsingh@mitsgwalior.in

Manish Dixit
Department of CSE and IT
MITS, Gwalior, India
dixitmits@mitsgwalior.in

Abstract: In recent scenario, network intrusions are increasing exponentially which in result compromised the protection of data. To protect data, intrusion detection systems are utilized to monitor intrusions on the network. Machine learning algorithms and classification algorithms are utilized to construct "Intrusion Detection Models" which can divide the network traffic into malicious or normal traffic. Huge amount of network traffic can collect the inappropriate and redundant features that can affect the performance of IDS in terms of accuracy and training time. In the proposed work, combinations of feature selection methods i.e. filter and wrapper are utilized to select best features by considering 41 attributes from the intrusion dataset (NSL_KDD Dataset) using Boolean AND operator. Classifications have been carried out on relevant attributes using three classifiers IBK, IB1, Random Committee. A comparative analysis is done with existing work and it is observed that the proposed work gives higher accuracy.

Keywords: Intrusion Detection System; NSL_KDD dataset; Classification Algorithms; Feature selection techniques; Boolean AND operator

1. INTRODUCTION

In recent scenario, network traffic is rapidly increasing and new vulnerabilities are found. Intrusion detection is extremely necessary to stop the attackers from interrupting or misusing the computer system. Intrusions are actions that can ruin the confidentiality, integrity and availability of computer system. An Intrusion detection system is a system that monitors network traffic and analyzes traffic for possible attacks generated from outside and within the organization and also for misuse of the system [1]. Intrusion detection system is classified into two types: Network based IDS (NIDS) and Host based IDS (HIDS). NIDS control all the traffic through the network section and when abnormal activity is discovered, it notifies system administrator. HIDS applications running on individual hosts on the network and alert the system user after detection of abnormal activity. The most common approaches to Intrusion Detection are Anomaly detection and Signature detection. Anomaly detection defines a collection of rules that rules can be used to decide that given behavior is that of an intruder. Rules are built to discover deviation from previous usage patterns. Signature detection monitors the network traffic and analyzes it against predefined attacks. When an attack is detected, an alarm is generated.

2. RELATED WORK

Raman Singh et. al [12] proposed a work for performance analysis of various classifiers after reducing a number of features . They analyzes the performance of various classification techniques and feature selection techniques in terms of accuracy, number of features, time taken to build model, TPR, FPR. Filtered subset evaluation gives

less computational time. This technique reduces features by 82.93% and gives better accuracy. CFS subset evaluator was the second best for network dataset. This technique decreases features by 75.61%.

P. Jongsuebsuk et. al. [13] proposed a Fuzzy Genetic algorithm for intrusion detection system that establish known attack types with highest accuracy and lowest false positive rate. Reduced features will be faster and uses less memory consumption than using all the features. This algorithm can also identify new attack types with high accuracy.

Ayman I. Madbouly et. al. [14] proposed a relevant feature selection model to select best features set that can be applied to design a lightweight intrusion detection system. Seven totally different feature selection methods were used to select and rank relevant features. The proposed model has four different phases, pre-processing of data, best selection of classifier, reduction of feature, and best feature selection. The result indicates that some features have no relevance to discover any attack type. Some features are useful to discover all types of attack. It has been proposed that a set of best 11 features find to be best against the full 41- features set. This model produces high detection rates as well as speed up the detection process.

Balakrishnan et. al [11] proposed a new feature selection model which is based on Information Gain Ratio. The proposed feature selection selects only the important features that facilitate in decreasing the classification time. The main advantage of the proposed IDS reduces the false positives and computation time.

Tanya Garg et. al. [7] proposed a model using different filter based feature selection models for reducing number of features. Ten classification algorithms have been carried out for further analysis. The combinations of six reduced attribute sets have been finalized using Boolean AND operator. Combination of Symmetric and Gain Ratio performs best.

3. DATASET DESCRIPTION

NSL_KDD is the improved version of KDDcup99 dataset. It removes all the duplicate records present in KDDcup99. It contains all important records of the complete KDD dataset. The training and testing dataset are mentioned in TABLE I.

TABLE I. Description of dataset

Class Type	Instances in KDDTrain	Instances in KDDTest
Normal	67343	9711
Dos	45927	745
Probe	11656	2421
U2R	52	200
R2L	995	2754
Total	125973	22544

4. PROPOSED WORK

In the proposed framework, the subset of best features has been achieved by reducing NSL-KDD [10] dataset of 41 features for intrusion detection. The features are selected by using the combination of filter method and wrapper method. The Boolean AND operation is performed on the selected features. The classifiers are utilized on the set of final features and the analysis is done according to the classification rate .The proposed framework is shown in Fig. 1 which has two phases:

Phase 1: Best Feature set selection using proposed Hybridized Combinational Feature Selection Framework (HCFSF)

Phase 2: Building up classification model using reduced feature set.

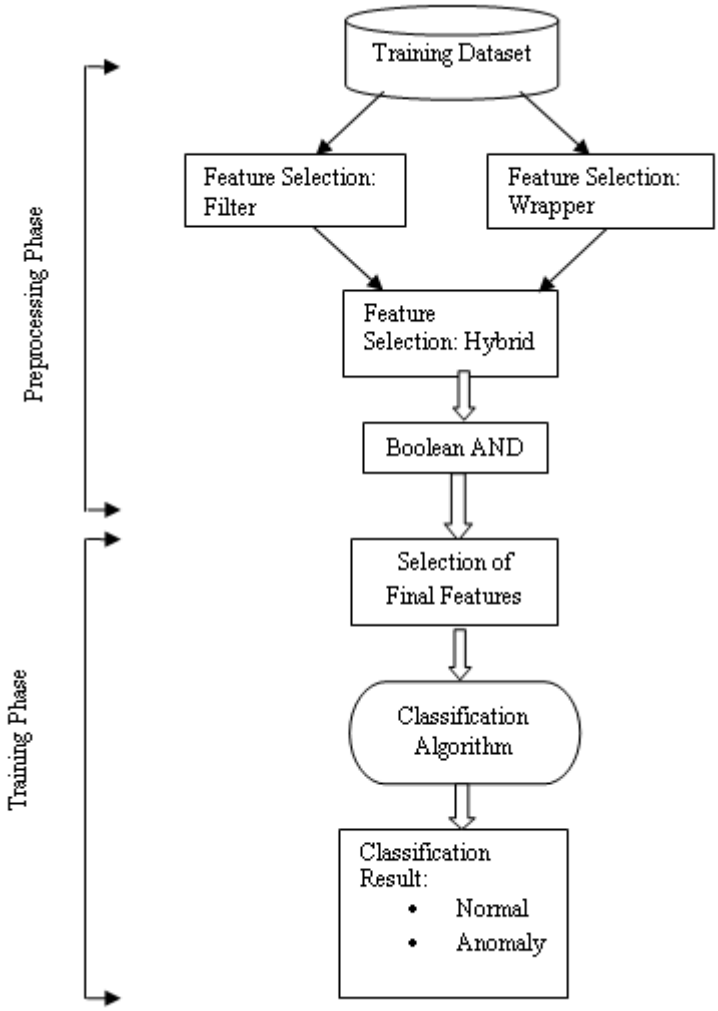

Fig. 1 The Proposed Framework

4.1. Feature Selection Strategy

Feature selection, in data mining, is the process of selecting appropriate attributes from the large set of attributes for the chosen dataset.

Feature selection is divided in two categories: Filter and Wrapper. In filter methodology, all the features are scored and ranked according to certain criteria. The features with the best rankings are selected and the low score features are eliminated.. Filtering methods are quick but free from the classifier but ignore the feature dependencies. They are easily adaptable to very large dataset. Therefore, the selection of these features are performed only once and after that different classifications can be performed. The disadvantage of filter models is that they ignore the interaction with the classifier and every feature is taken into account thus discarding feature dependencies. Additionally, the threshold point is not determined for rankings to select only the desired features but also ignore the noise [5]. Filter methodology shown in Fig. 2 [5].

The Wrapper method is used to calculate attribute weights by using the classification model to measure the performance. The wrapper method interacts with the classifier, shows feature dependency, offers good accuracy of the classification and reduces computational cost. The disadvantage with wrapper method is that they are time consuming and the results depend on the classifiers and works well with the smaller dataset and they are also computationally expensive. The wrapper model is shown in Fig. 3 [5].

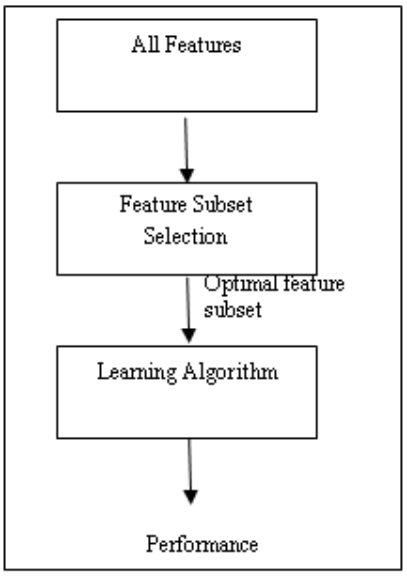

Fig. 2: The Feature Filter Method

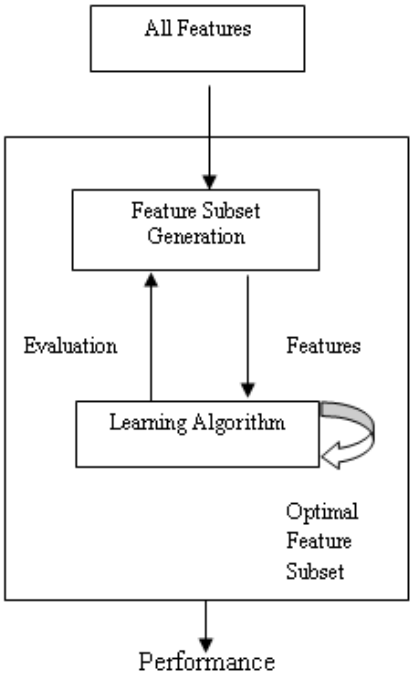

Fig. 3: The Feature Wrapper Method

The proposed work is focused on the combination of both filter and wrapper method i.e. Hybrid method. Both feature selection methods have some drawbacks. Then combination of both performs best and gives better result in terms of accuracy, computational time, classification rate. In our research work, four ranking based feature selection techniques are used:

4.1.1. Information Gain (IG): Information Gain is the filter based feature selection method based on the concept of entropy. Entropy is a measurement of purity or impurity of a variable . The disadvantage of Information gain is it chooses different characteristic that having large unique values over the features having fewer unique values, although they are more instructive [6].

4.1.2. Gain Ratio (GR): Gain ratio is the ratio between information gain to the real value. Gain ratio is the modification of Information gain to reduce its biasness. It takes many branches for selecting the required attribute. It is based on the information provided by intrinsic attributes. Intrinsic information is the information used to determine which branch belongs to which instance. In this method the intrinsic information gets larger when the value of attribute is decrease. Gain ratio selects an attribute with low intrinsic value. The attribute which has highest gain ratio will be chosen as the splitting attribute [7].

*4.1.3. Symmetrical Uncertainty***:** To calculate the fitness of features between feature and the target class, symmetric uncertainty is used. The feature with high symmetric uncertainty gets high importance [8]. It is a correlation based Feature Selection approach that measures the quality of a feature in a set using a hypothesis – "Good feature subsets contain features highly correlated with the class, yet uncorrelated to each other". This Feature selection technique used to measure the degree of association between the distinct attributes. It is also based on the idea of entropy and could be a symmetric measure to measure correlation between set of features.

4.1.4. OneR: It is a rule based algorithm that creates rules for ranking and selecting the features. OneR creates rules and tests one by one attribute and branch for each value of that attribute [7].

4.1.5. Chi-Square: This method is based on the statistical theory. It measures the independence and power of relationship between two attributes. The prediction of an attribute value is measured using observed and expected values of attributes.

Three wrapper based feature selection techniques are used in the proposed work:

*4.1.6. CfsSubsetEval***:** The value of defined attributes set is expressed not only by individual predictive ability of each feature but also the degree of redundancy between them. The features with low intercorrelation and extremely correlate with the class are preferred. A feature is valuable if it is strongly correlated with the class but not much correlated with other features of the class [9].

*4.1.7. ConsistencySubsetEval***:** It evaluates the useful set of attributes by the level of consistency within the class values once the training instances are projected onto the set of attributes. The Consistency of any set of attributes should be greater than that of the total set of attributes. To use this subset evaluator with a Random or Exhaustive search, the consistency of the small subset can be equal to the total set of attributes [15].

4.1.8. FilteredSubsetEval: It has class for running an arbitrary subset evaluator on data that has been passed through an arbitrary filter. Like the evaluator, the structure of the filter is based exclusively on the training data.

4.2. Building up Classification Models

Classification algorithms or machine learning algorithm are used to classify network traffic when they are used on dataset.

In the proposed work, there are 76 classifiers that are suitable for preferred dataset. But we have chosen best classifiers IB1, IBK and Random Committee that are used in existing work [7]. They give better result on the basis of their performance. These three classifiers are:

4.2.1. IBK : This classifier is known as the lazy learner because it takes more time for computation. This classifier takes large amount of time for classification as a result of each example that must be classified should be compared to the each example in training dataset [4].

4.2.2. IB1: It is also called as Nearest-neighboring classifier. This classifier uses Euclidean distance to find the training instance closed to the determined test instance, and realize the same class as this training instance. If several instances have the same distance to the test instance, the first instance is used [15].

4.2.3. Random Committee: This classifier comes under meta classifier. This classifier finds the best set of attributes to train the base classifier with these parameters [3], This trained base classifier are going to be used for another predictions. The final prediction is an average of the prediction generated by the individual base classifier.

5. EXPERIMENTAL RESULTS

All the evaluations have been done on the reduced features set after feature selection of the NSL-KDD Dataset using three classification algorithms. These classifications were evaluated using version 3.6.13 of Weka data mining tool. Evaluations have been accomplished to compare the performance of combinations of six filter based feature selection techniques: Info Gain (IG), Gain Ratio (GR), , OneR, Chi-square, Symmetrical Uncertain and Filtered attribute evaluator and three wrapper based feature selection techniques: CfsSubsetEval (CFS), ConsistencySubsetEval and FilteredSubsetEval. The first 20 features selected by different combination of two and three feature selection techniques have been combined by using Boolean AND operator. The performance of resultant set has been carried out using the classification algorithms that are used in the existing work. The top most hybrid combinations of proposed work are listed in the TABLE III. The reduced sets of features by top most combinations are mentioned in TABLE II. And the comparative study of existing work and proposed work is mentioned in TABLE IV.

In the proposed work, a hybridized combinational framework for feature selection is being proposed in TABLE III. From the TABLE III, it can be observed that the combination of Gain Ratio and consistency subset evaluator are the worst feature selector. They gives worst performance in all aspect showing highest FPR , error and lowest classification rate, recall, precision, ROC and kappa.

A comparative study of existing work [7] and proposed work is mentioned in TABLE IV and from the table it can be observed that the feature selection techniques of the proposed work gives better result in aspect of accuracy with the same classifier used in existing work. And it also can be observed that CFS Subset Evaluator (wrapper method) gives better result when combining with filter methods.

TABLE II. List of features for top combinations

Name of Technique	Features Selected
Symmetric+ GR+ CFS	2,3,4,5,6,12,25,29,30,35,38,39
Symmetric+ GR+ Consistency	38,39
Symmetric+ GR+ Filtered	2,3,4,5,6,12,25,29,30,38,39
OneR+ Symmetric+ CFS	3,4,5,6,12,23,25,29,30,35,36,37,38,39
OneR+ Symmetric+ Consistency	3,5,6,35,38,39
OneR+ Symmetric+ Filter	3,4,5,6,12,23,25,29,30,35,36,38,39
Symmetric+ IG+ CFS	3,4,5,6,12,23,25,29,30,35,36,37,38,39
Symmetric+ IG+ Consistency	3,5,6,35,38,39
Symmetric+ IG+ filter	3,4,5,6,12,23,25,29,30,35,36,37
OneR+ Symmetric+ GR+ CFS	2,3,4,5,6,12,25,29,30,35,38,39
OneR+ Symmetric+ GR+ Consistency	38,39
OneR+ Symmetric+ GR+ Filter	2,3,4,5,6,12,25,29,30,38,39
GR+ IG+ CFS	3,4,5,6,12,25,29,30,35,38,39
GR+ IG+ Filter	3,4,5,6,12,25,29,30,38,39

GR+ IG+ Consistency	38,39

TABLE III. Top hybrid combinations of feature selection techniques

Name of Technique	Classifier	Training Time	Accuracy	FPR	ROC	Recall	Precision	Error	Kappa
Symmetric+ GR+ CFS	IBK	0.08 sec	99.2062	0	1	0.992	0.994	0.001	0.9869
Symmetric+ GR+ Consistency	IBK	0.05 sec	82.1089	0.194	0.853	0.821	0.789	0.0249	0.6665
Symmetric+ GR+ Filtered	IBK	0.05 sec	99.1212	0.001	1	0.991	0.993	0.0011	0.9855
OneR+ Symmetric+ CFS	Random Committee	11.24 sec	99.9738	0	1	1	1	0	0.9996
OneR+ Symmetric+ Consistency	Random Committee	3 sec	98.7196	0.001	1	0.987	0.989	0.0016	0.9788
OneR+ Symmetric+ filter	Random committee	2.87 sec	99.2165	0	1	0.992	0.994	0.0009	0.987
Symmetric+ IG+ CFS	IBK	0.08 sec	99.973	0	1	1	1	0	0.9996
Symmetric+ IG+ Consistency	IBK	0.05 sec	98.7196	0.001	1	0.987	0.989	0.0016	0.9788
Symmetric+ IG+ filter	IBK	0.08 sec	99.9722	0	1	1	1	0	0.9995
OneR+ Symmetric+ GR+ CFS	Random Committee	9.63 sec	99.2062	0	1	0.992	0.994	0.001	0.9869
OneR+ Symmetric+ GR+ Consistency	Random Committee	1.95 sec	82.1089	0,194	0.853	0.821	0.789	0.0249	0.6665
OneR+ Symmetric+ GR+ Filter	Random Committee	9.71 sec	99.1212	0.001	1	0.991	0.993	0.0011	0.9855
Symmetric+ GR+ CFS	IB1	0.06 sec	99.199	0	0.996	0.992	0.993	0.0007	0.9867
Symmetric+ GR+ Consistency	IB1	0.05 sec	81.3214	0,191	0.811	0.813	0.778	0.156	0.6546
Symmetric+ GR+ Filter	IB1	0.09 sec	99.1038	0.001	0.995	0.991	0.992	0.0007	0.9852
OneR+ Symmetric+ CFS	IB1	0.06 sec	99.9643	0	1	1	1	0	0.9994
OneR+ Symmetric+ Consistency	IB1	0.03 sec	98.564	0.001	0.992	0.986	0.987	0.0012	0.9762
OneR+ Symmetric+ Filter	IB1	0.06 sec	99.2165	0	0.996	0.992	0.994	0.0007	0.987
GR+ IG+ CFS	IBK	0.06 sec	99.2062	0	1	0.992	0.994	0.001	0.9869
GR+ IG+ Consistency	IBK	0.06 sec	82.1089	0.194	0.853	0.821	0.789	0.0249	0.6665
GR+ IG+ Filter	IBK	0.06 sec	99.1212	0.001	1	0.991	0.993	0.0011	0.9855
Symmetric+ IG+ CFS	IB1	0.17 sec	99.9643	0	1	1	1	0	0.9994
Symmetric+ IG+ Consistency	IB1	0.09 sec	98.564	0.001	0.992	0.986	0.987	0.0012	0.9762
Symmetric+ IG+ Filter	IB1	0.08 sec	99.9587	0	1	1	1	0	0.9993

TABLE IV. Comparative study with existing work and proposed work (HCFSF)

	Existing Work [7]		**Proposed Work (HCFSF)**	
Classification Technique	Name of Technique	Accuracy	Name of Technique	Accuracy
IBK	Symmetric+ GR	98.5	Symmetric+ GR+ CFS	99.2062
Random Committee	OneR+ Symmetric	97.34	OneR+ Symmetric+ CFS	99.9738
IBK	Symmetric+ IG	96.2	Symmetric+ IG+ CFS	99.973
Random Committee	OneR+ Symmetric+ GR	95.4	OneR+ Symmetric+ GR+CFS	99.2062
IB1	Symmetric+ GR	96	Symmetric+ GR+ CFS	99.199
IB1	OneR+ Symmetric	94.6	OneR+ Symmetric+ CFS	99.9643
IBK	GR+IG	94	GR+IG+CFS	99.2062

6. CONCLUSION

In the proposed work, various combinations of filter based and wrapper based feature selection techniques are utilized. The classification algorithms have been carried out on the reduced attributes. All the analysis in the proposed work have been accomplished using Hybrid Feature selection model. A Hybridized combinational feature selection framework combines the reduced feature sets resulted by two and three feature selection techniques using Boolean AND operation. In the proposed work, it is observed that when we combine CFS with existing technique with the same classifier, it will give better result in comparison with existing work. And it is also observed that the combination of Gain Ratio and Consistency gives worst performance in all aspect.

References

[1] https://www.sans.org/readingroom/whitepapers/detection/intrusion-detection-systems-definition-challenges-343.

[2] Dash R.," Selection of the Best Classifier from Different Datasets Using WEKA", HERT, Vo1.2 Issue 3, March 2013.

[3] Garg T. and Khurana S. S.," Comparison of Classification Techniques for Intrusion Detection Dataset Using WEKA", IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), May 2014.

[4] Panda M. and Patra M., '"Ensembling Rule Based Classifiers for Detecting Network Intrusions", IEEE Conference on Advances in Recent Technologies in Communication and Computing, 2009.

[5] Kumari B., Swarnkar T.," Filter versus Wrapper Feature Subset Selection in Large Dimensionality Micro array: A Review", IJCSIT 2011.

[6] Witten I. H., Frank. E., Practical Machine Learning Tools and Techniques (2nd ed.), 2012.

[7] Garg T., Kumar Y.," Combinational Feature Selection Approach for Network Intrusion Detection System", 2014 International Conference on Parallel, Distributed and Grid Computing.

[8] Singh B. , Kushwaha N., Vyas O. P.," A Feature Subset Selection Technique for High Dimensional Data Using Symmetric Uncertainty", Journal of Data Analysis and Information Processing, 2014.

[9] Doshi M. , Dr. Chaturvedi S. K.," Correlation Based Feature Selection (CFS) Technique to Predict Student Performance", International Journal of Computer Networks & Communications (IJCNC) Vol.6, No.3, May 2014.

[10] Nsl-kdd data set for network-based intrusion detection systems. Available on: " https://github.com/defcom17/NSL_KDD"

[11] Balakrishnan S., Venkatalakshmi K and Kannan A, "Intrusion Detection System Using Feature Selection and Classification Technique", International Journal of Computer Science and Application (IJCSA), Nov 2014.

[12] Singh R., Kumar H. and Singla R. K., "Analysis of Feature Selection Techniques for Network Traffic Dataset", International Conference on Machine Intelligence Research and Advancement, IEEE Dec. 2013.

[13] Jongsuebsuk P. and Wattanapongsakorn N., "Network Intrusion Detection with Fuzzy Genetic algorithm for Unknown Attacks", Information Networking (ICOIN), International Conference, IEEE,2013.

[14] Madbouly A I, Gody A M, Tamer and Barakat M ," Relevant Feature Selection Model Using Data Mining for Intrusion Detection System", International Journal of Engineering Trends and Technology(IJETT), March 2014.

[15] http://weka.sourceforge.net/doc.stable/weka/attributeSelection/ConsistencySubsetEval.html.