Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods

Zahra Karimi
Islamic Azad University
Tehran North Branch
Dept. of Computer Engineering
Tehran, Iran

Mohammad Mansour Riahi Kashani
Islamic Azad University
Tehran North Branch
Dept. of Computer Engineering
Tehran, Iran

Ali Harounabadi
Islamic Azad University
Central Tehran Branch
Dept. of Computer Engineering
Tehran, Iran

ABSTRACT
Intrusion detection is a crucial part for security of information systems. Most intrusion detection systems use all features in their databases while some of these features may be irrelevant or redundant and they do not contribute to the process of intrusion detection. Therefore, different feature ranking and feature selection techniques are proposed. In this paper, hybrid feature selection methods are used to select and rank reliable features and eliminate irrelevant and useless features to have a more accurate and reliable intrusion detection process. Due to the low cost and low accuracy of filtering methods, a combination of these methods could possibly improve their accuracy by a reasonable cost and create a balance between them. In the first phase, two subsets of reliable features are created by application of information gain and symmetrical uncertainty filtering methods. In the second phase, the two subsets are merged, weighted and ranked to extract the most important features. This feature ranking which is done by the combination of two filtering methods, leads to higher the accuracy of intrusion detection. KDD99 standard dataset for intrusion detection is used for experiments. The better detection rate obtained in proposed method is shown by comparing it with other feature selection methods that are applied on the same dataset.

Keywords
Intrusion Detection, Feature Selection, Filtering, KDD99 Dataset

1. INTRODUCTION
Feature selection is a pre-processing technique that finds a minimum subset of features that captures the relevant properties of a dataset to enable adequate classification [1]. Given that no loss of relevant information is incurred with a reduction in the original feature space, feature selection has been widely used. Feature selection has been considered in many classification problems [2], and it has been used in various application domains [3], [4]. Feature selection techniques are very useful for improving the performance of learning algorithms [5]. For this reason, the strengths and weaknesses of feature selection techniques are traditionally assessed in terms of the classification performance from models built with a subset of the original features.

Therefore, the hybrid feature selection method was used in this paper to select and rank reliable features and eliminate irrelevant and useless features to have a more accurate and reliable intrusion detection process and to eliminate the possibility that a single feature selection approach will result to some biased results. Therefore, combining them is a good and reasonable choice.

In the first phase, two subsets of reliable features are created by application of information gain and symmetrical uncertainty filtering methods. In the second phase, the two subsets are merged, weighted and ranked to extract the most important features. This feature ranking which is done by the combination of two filtering methods, leads to higher the accuracy of intrusion detection.

The definition of Feature selection, Filtering Methods, Intrusion detection, and Naïve Bayes classifier which are used in proposed method, is presented in section 2, 3, 4 and 5. In section 6, proposed method and the phases involved in the feature selection process is described. In section 7, the performance of the proposed method is tested on KDD99 dataset. Conclusions are given in section8.

2. FEATURE SELECTION
In order to make IDS more efficient, reducing the data dimensions and complexity have been used as simplifying features. Feature selection can reduce both the data and the computational complexity. It can also get more efficient and find out the useful feature subsets. It is the process of choosing a subset of original features so that the feature space is optimally reduced to evaluation criterion. The raw data collected is usually large, so it is desired to select a subset of data by creating feature vectors that Feature subset selection is the process of identifying and removing much of the redundant and irrelevant information possible. This results in the reduction of dimensionality of the data and thereby makes the learning algorithms run in a faster and more efficient manner.

The feature selection techniques are generally mainly divided into two categories, filter and wrapper [6]. Filter method operates without engaging any information of induction algorithm. By using some prior knowledge such as feature should have strong correlation with the target class or feature should be uncorrelated to each other, filter method selects the best subset of features Alternatively, wrapper method employs a predetermined induction algorithm to find a subset of features with the highest evaluation by searching through the space of feature subsets and evaluating quality of selected features. The process of feature selection acts like "wrapped around" an induction algorithm since wrapper approach includes a specific induction algorithm to optimize feature selection; it often provides a better classification accuracy result than that of filter approach. However, wrapper method is more time consuming than filter method due to it is strongly coupled with an induction algorithm with repeatedly calling the algorithm to evaluate the performance of each subset of features. It thus becomes unpractical to apply a wrapper method to select features from a large data set that contains

numerous features and instances [7]. Furthermore, wrapper approach is required to re-execute its induction algorithm for selecting features from data set while the algorithm is replaced with a dissimilar one. It is less independent of any induction algorithms than filter is.

3. FILTERING METHODS

There are seven filter-based feature ranking techniques that are described in below: The first six are commonly used in the literature (chi-squared statistic ($\chi 2$), Information Gain (IG), Gain Ratio (GR), two versions of Relief (RF and RFW) and Symmetric Uncertainty (SU), while the last, Signal-to-noise (S2N), is less well known. $\chi 2$, IG, GR,RF, RFW and SU are available in the Weka data mining tool [8]. $\chi 2$, IG, GR and SU utilize the method to discretize continuous attributes, and all four methods are bivariate, considering the relationship between each attribute and the class, excluding the other independent variables [9].

3.1 Information Gain

Information Gain (IG) is a commonly used measure in the fields of information theory and machine learning. IG measures the number of bits of information gained about the class prediction when using a given feature to assist that prediction [10]. For each feature, a score is obtained based on how much more information about the class is gained when using that feature. The information gain of feature X is shown in equation 1:

$$IG(X) = H(Y) - H(Y|X) \quad (1)$$

Where H(Y) and H(Y |X) are the entropy of Y and the conditional entropy of Y given X, respectively. The level of a feature's significance is thus determined by how great is the decrease in entropy of the class when considered with the corresponding feature individually.

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.

3.2 Symmetrical Uncertainty

Symmetric Uncertainty (SU) is a correlation measure between the features and the class [8]. And it is obtained by equation 2:

$$SU = \frac{H(X)+H(Y)-H(X|Y)}{H(X)+H(Y)} \quad (2)$$

where H(X) and H(Y) are the entropies based on the probability associated with each feature and class value respectively and H(X,Y), the joint probabilities of all combinations of values of X and Y .

A weakness of the IG criterion is that it is biased toward features with fewer values.

4. INTRUSION DETECTION

An intrusion detection system (IDS) can be a device or software application that monitors the network or system activities for malicious attacks or policy violations and reports it to a Management Station [11]. IDS are considered to provide dynamic defense mechanisms to various network security threats. IDS can be divided into two types as network based and host based. Network intrusion detection system (NIDS) detects intrusions by continuously monitoring network traffic by connecting to network hub or switch which is configured for port mirroring, or network tap. NIDS uses sensors to capture all network traffic and to monitor individual packets to identify whether it is normal or attack. An example of a NIDS is Snort [12]. Host-based intrusion detection system (HIDS) uses agent as a sensor on a host that identifies intrusions by analyzing system calls, application logs, file-system modifications (binaries, password files, etc.) and other host activities and state. OSSEC is an example for Host based intrusion detection system [12].

Passive systems are called as Intrusion Detection Systems and reactive systems are known as Intrusion Prevention Systems. IDS detect malicious activity, from a set of log records and alert the user. IPS auto-responds to the suspicious activity by resetting the connection or by reprogramming the firewall to block network traffic from the suspected malicious source. Based on the methodology adopted to identify intrusions, IDS could be classified as: anomaly detection and misuse detection. In anomaly detection, normal user behavior is developed. The anomaly detector monitors incoming packets and check for normal behavior. If it is deviating then it is considered as abnormal or attack. In misuse detection, abnormal behavior is modeled [13]. The misuse detector monitors network segments and check for abnormality. Misuse detector has higher accuracy when compared to anomaly detector because modeling normal behavior is difficult. Commercial IDS are mostly based on misuse detection [13]. The log records usually contain a large number of features which make the task of an Intrusion Detection System very difficult. Hence important features can be derived using some feature reduction algorithm and used for classification of data as normal or attack.

4.1 Intrusion Detection Dataset

The data set used for the entire course of research is the DARPA KDD99 benchmark data set, also known as "DARPA Intrusion Detection Evaluation data set". It includes three independent sets: whole KDD, 10% KDD, and corrected KDD. In the experiments, 10% KDD and corrected KDD are taken as training and testing set, respectively. The training set contains a total of 22 training attack types, with an additional 17 types in the testing set only. Totally 39 attack types are included and are fall into four main classes, Denial of Service (DOS), Probe, User to Root (U2R), and Remote to Local (R2L).Both training and testing sets are made up of a large number of network traffic connections and each one is represented with 41 features plus a label of either normal or a type of attack. The training set includes 494,020 connections that are distributed as 97,277 normal connections, 391,458 DOS attacks, 4,107 Probe attacks, 52 U2R attacks, and 1,126 R2L attacks. The testing set has 311,029 connections. It is made up of 60,593 normal connections, 229,853 DOS attacks, 4,166 Probe attacks, 228 U2R attacks, and 16,189 R2L attacks [14].

5. NAIVE BAYES

This classifier is based on the elementary Bayes' Theorem. It can achieve relatively good performance on classification tasks [36]. Naive Bayes classifier greatly simplifies learning by assuming that features are independent given the class variable. More formally, this classifier is defined by equation 3:

$$F_i(X) = \prod_{j=1}^{N} P(X_j|C_i)P(C_i) \quad (3)$$

Where X = ( ) denotes a feature vector and Cj, j = 1, 2, ..., N, denote possible Class labels.

The training phase for learning a classifier consists of estimating conditional probabilities P (XJ | Ci) and prior probabilities. Here, are estimated by counting the training examples that fall into class ci and then dividing the resulting count by the size of the training set. Similarly, conditional

probabilities are estimated by simply observing the frequency distribution of feature XJ within the training subset that is labeled as class ci. To classify a class-unknown test vector, the posterior probability of each class is calculated, given the feature values present in the test vector; and the test vector is assigned to the class that is of the highest probability [15].

6. PROPOSED METHOD

Input: S ($f_1, f_2, \ldots, f_N, C$) // a testing data set for attack
δ // a predefined Rank
Ś ($f_1, f_2, \ldots, f_M, C$) // a selection subset

Output: S_{Best} ($f_1, f_2, \ldots, f_L, C$) // an optimal subset

Begin

For i = 1 to N do
Begin
Calculate IG for f_i ;
Calculate Rank ($IG_{i,c}$) for f_i;
If (Rank ($IG_{i,c}$) ≤ δ)
Append f_i to $\acute{S}_{List1}$
End;

For i = 1 to N do
Begin
Calculate SU for f_i;
Calculate Rank ($SU_{i,c}$) for f_i;
If (Rank ($SU_{i,c}$) ≤ δ)
Append f_i to $\acute{S}_{List2}$
End;

$\acute{S}_{List}$ = Combine ($\acute{S}_{List1}$,$\acute{S}_{List2}$)

For i = 1 to M do
Begin
$IG - SU_{i,c}$ = AVG (Rank ($IG_{i,c}$), Rank ($SU_{i,c}$))
If ($IG - SU_{i,c}$) ≤ δ)
Append f_i to S_{Best}
End;

For i = 1 to L do
Calculate Weight ($IG - SU_{i,c}$) for f_i;

Figure 1: proposed method

According to the topics discussed in this paper, a method was proposed for selection of the effective features by using a combination of two algorithms of information gain and symmetric uncertainty that consists of this aspect "how the features that most correlated with the class of attack should be selected?" To answer this question, an attempt has been made to use the concept of features ranking that its default value is decided (selected) by the user. As it can be seen in Figure 1, with given a data set with a number of input features and a target class, the proposed method in first and second stage by using of two algorithms of information gain and symmetric uncertainty, first calculates the mutual information between features and class, Then the features are classified in descending order based on the degree of importance of relevance with the class, and the relevance with the class of those features that their degree of importance is higher than threshold level is kept, i.e. the deleted features are completely irrelevant to the class and the remaining ones (the rest of the features) are predictable.

At third stage, the remaining features of the first and second stages are combined and the average importance factor of each feature is calculated. Those features that their average importance factor is higher than the threshold level are kept. Finally, the weight of the remaining features is calculated based on the average importance factor and by using of the weighted mean. In addition, the final rank of the remaining features is defined based on the weight, which means the selected features are the most "significant features" that restrain indispensable information of the original feature space.

7. EMPIRICAL STUDY

7.1 Experimental Conditions

In order to evaluate the performance of proposed feature selection algorithm on data sets, four representative feature selection algorithms, IG,SU,CFS and FSSCP, built on the top of symmetric uncertainty are chosen. CFS method [16] uses a correlation-based heuristic search algorithm to evaluate the worth of subsets of features. It considers good feature subsets contain features that are highly correlated with the class, yet uncorrelated with one another. The heuristic algorithm measures the merit of feature subsets from pair wise feature correlations and then the subset with the highest merit found during the search is reported. Rather than scoring the worth of subsets of features of CFS approach, FSSCP method [14] uses symmetric uncertainty to evaluate the worth of features and then eliminate both irrelevant features with poor prediction ability to the class and redundant features that are inter correlated with one or more of the other features. After removing irrelevant and redundant features, the remaining ones contain indispensable information about the original feature space keeps. The collection features the strength of correlation between each pair of attributes will compute. The total amount of mutual information for each feature is acquired by adding all mutual information measures together that relate to that feature. In addition, the Naïve Bayes machine learning algorithm was applied to evaluate the detection accuracy on selected features for each feature selection algorithm.

7.2 Evaluation Parameters

In the experiments, the standard measurements such as detection rate (DR), false positive rate (FPR) were used for evaluation of the performance of intrusion detection tasks. The denotations of True Positive (TP), True Negatives (TN), False Positive (FP), and False Negative (FN) are defined as follows. Equations 4 and 5 describe DR and FPR, respectively [14].

True Positive (TP): The number of malicious records that are correctly identified.

True Negatives (TN): The number of legitimate records that are correctly classified.

False Positive (FP): The number of records that were incorrectly identified as attacks however in fact they are legitimate activities.

False Negative (FN): The number of records that were incorrectly classified as legitimate activities however in fact they are malicious.

$$DR = \frac{TP}{TP + FN} \quad (4)$$

$$FPR = \frac{FP}{TN + FP} \quad (5)$$

7.3 Experimental Results

The feature selection experiments were performed on the KDD99 data set. Four new sets of data are generated according to the normal class and four categories of attack (DOS, Probe, U2R and R2L). In each data set, records with the same attack category and all the normal records are included. For each data set, we run proposed approach and the other feature selection algorithms CFS and FSSCP, IG, SU, and record these selected features from each algorithm. During all experiments the threshold is considered by the user. Having finished the feature selection procedures, then the Naive Bayes machine learning algorithm was applied on each original full data set as well as each newly obtained data set that includes only those selected features from feature selection algorithms. By applying 10-fold cross-validation evaluation on each data set, the classification accuracies of this experimental dataset were obtained. Figure 2 show the results of feature selection of KDD99 dataset.

Figures 2 to 5 and 6 to 9 show the percentages of DRs and FPRs performed on four KDD99 data set using Naive Bayes algorithm, respectively.

Figures 10 and 11 show the average percentages of DRs and FPRs performed on four KDD99 data set using Naïve Bayes algorithm, respectively

For an intrusion detection task, abnormal activities are expected to be correctly identified and normal activities are anticipated not to be misclassified. Therefore, a higher DR and a lower FPR are desired.

Table 1.Selected Features in different methods on Normal-Dos dataset

IG-SU	IG	SU	FSSCP	CFS
5,2,23,36	5,23,3,36 ,24,2,33	6,12,5,3,24 ,32, 23, 37	1-6 ,12,23,24, 31,32, 37	5,6,11,12 ,31, 32

Table 2.Selected Features in different methods on Normal-Probe dataset

IG-SU	IG	SU	FSSCP	CFS
27,5,29,4	5,3,35,6, 27,23,37	25,29,27,4, 30, 5	1-4, 12,16,25, 27-29 ,30,40	40,5,25,27 ,29,30,37, 38,43

Table 3.Selected Features in different methods on Normal-U2R dataset

IG-SU	IG	SU	FSSCP	CFS
14,13,10, 17	3, 14,10,1 ,13, 33, 17	14,13,17, 10, 18, 29	1-3, 10, 16	1,14,17,29

Table 4.Selected Features in different methods on Normal-R2L dataset

IG-SU	IG	SU	FSSCP	CFS
3,10,5,33	5,3,6,33, 10, 36, 37	10,22,11,3, 33, 5, 6	1-5,10,22 , 26, 33	9,10,16

In tables 1 to 4, selected features in different methods for different attack datasets are presented. As it is clear from the tables, proposed method achieves higher degree of dimensionality reduction comparing to other methods. Moreover this reduction does not lead to deteriorate the classification performance because the precision of feature selection is improved by combining two filtering methods.

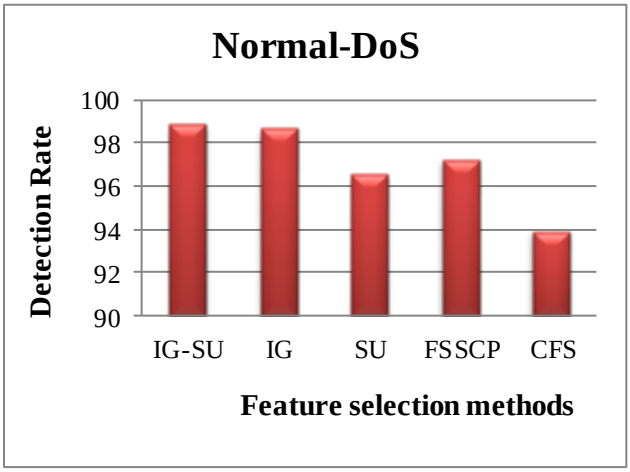

Figure 2: Detection Rate calculated for different methods in Normal-Dos dataset

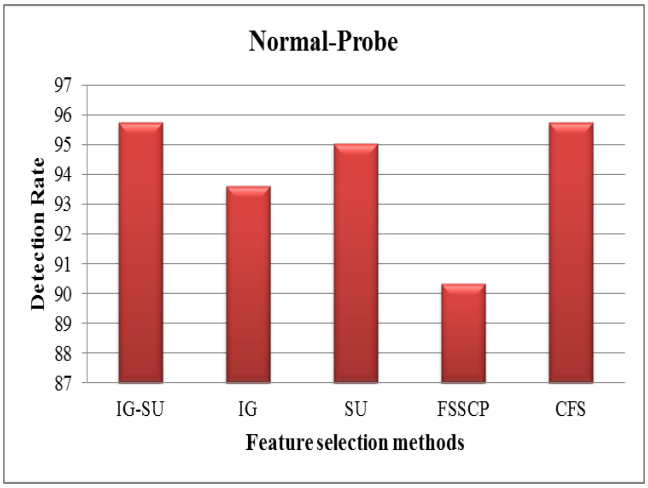

Figure 3: Detection Rate calculated for different methods in Normal-Probe dataset

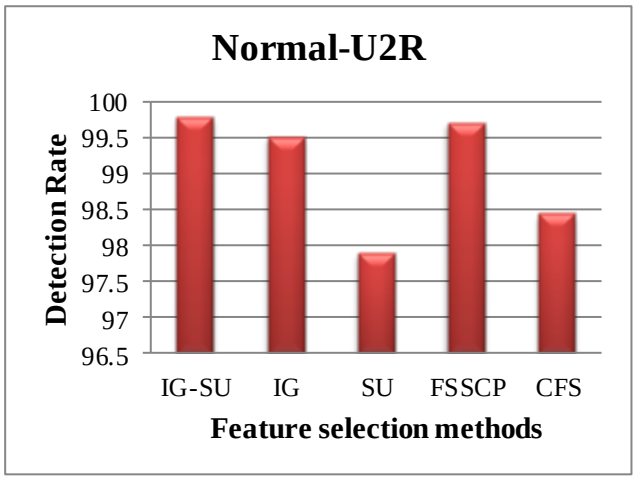

Figure 4: Detection Rate calculated for different methods in Normal-U2R dataset

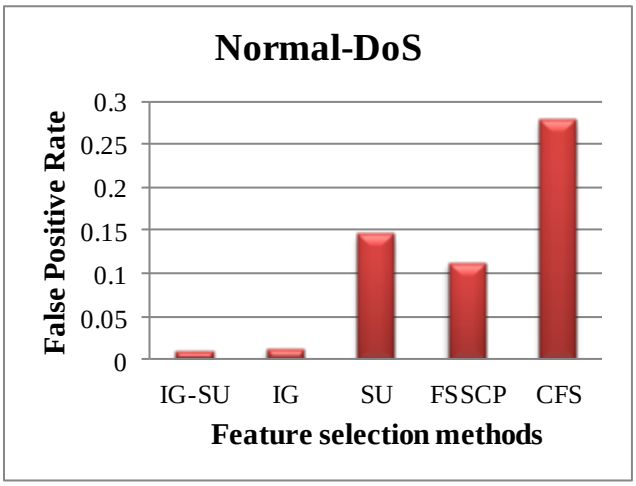

Figure 6: False Positive Rate calculated for different methods in Normal-Dos dataset

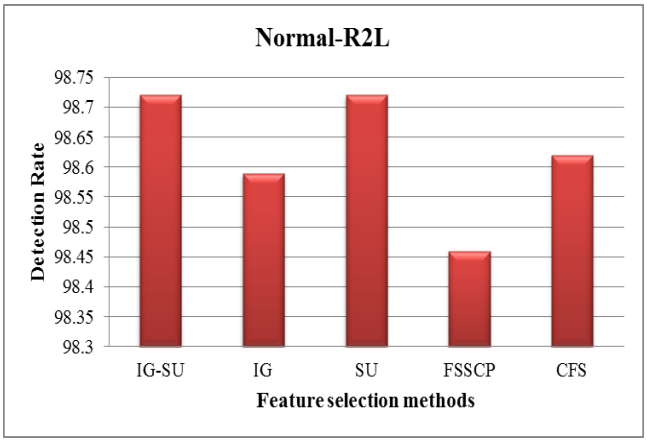

Figure 5: Detection Rate calculated for different methods in Normal-R2L dataset

In the figures 2 to 5, detection rate parameter in different methods for four attack datasets (Normal-Dos, Normal-Probe, Normal-U2R and Normal-R2L) are shown. From the figures above, we observe that in all the datasets, proposed method (combination of Information Gain and Symmetrical Uncertainty) achieves higher detection rate in comparison with other methods. This shows that the combination of two filtering methods, leads to improve the performance of classification.

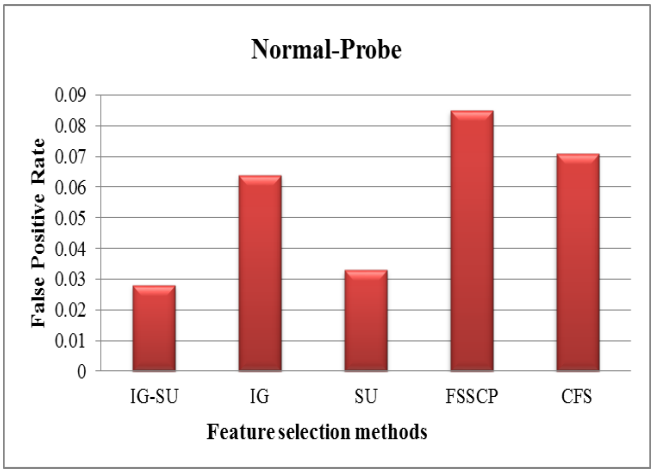

Figure 7: False Positive Rate calculated for different methods in Normal-Probe dataset

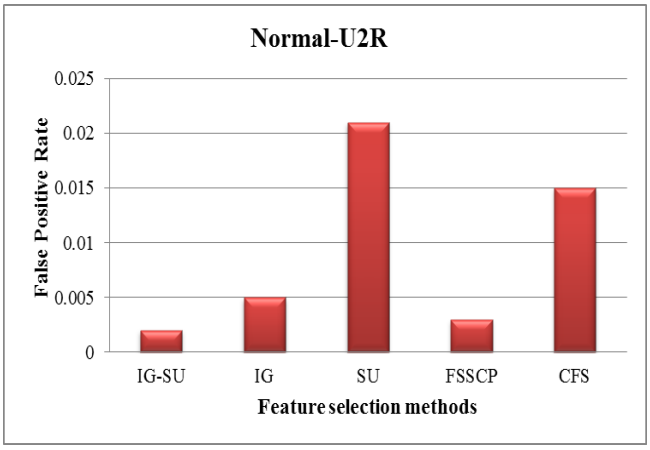

Figure 8: False Positive Rate calculated for different methods in Normal-U2R dataset

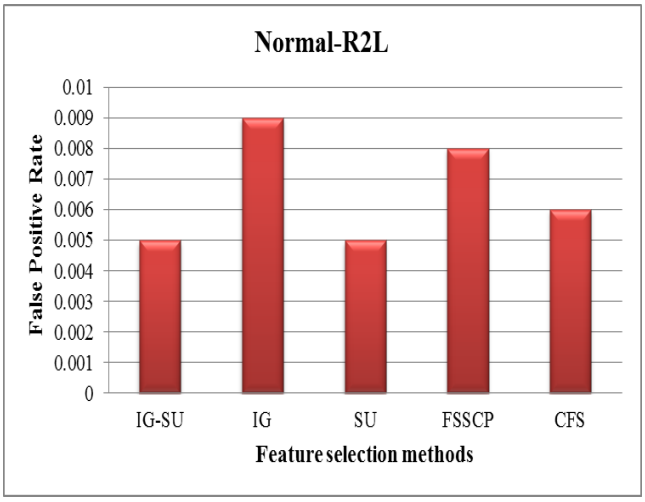

Figure 9: False Positive Rate calculated for different methods in Normal-R2L dataset

In figures 6 to 9, false positive rate parameters in different methods for four attack datasets (Normal-Dos, Normal-Probe, Normal-U2R and Normal-R2L) are shown. False Positive Rate parameter is the rate of records that were incorrectly identified as attacks however in fact they are legitimate activities. Hence, the lower value for this parameter is desirable. From the figures above, were observe that in all the datasets, proposed method (combination of Information Gain and Symmetrical Uncertainty) achieves lower false positive rate in comparison with other methods. This shows that the combination of two filtering methods, leads to lower the number of incorrectly classified records and improve performance of classification.

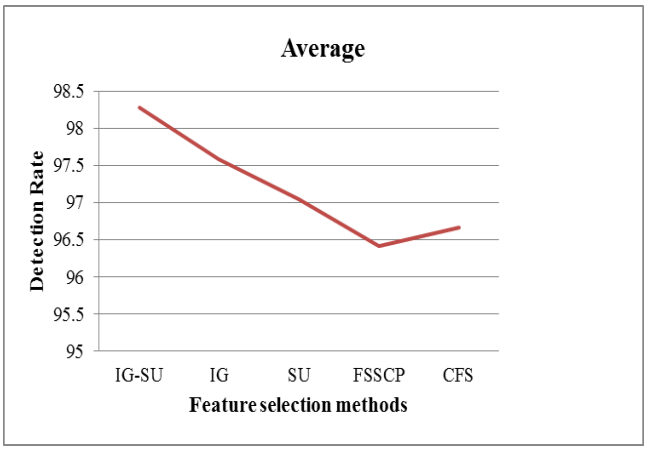

Figure 10: Average Detection Rate calculated for 4 attack datasets

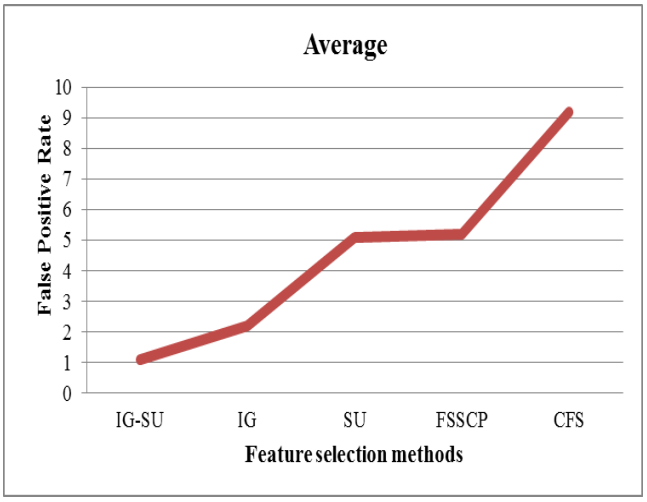

Figure 11: Average False Positive Rate calculated for 4 attack datasets

From the results shown in Figure 10, we observe that our approach achieves higher classification accuracy in comparison with the outcomes of IG, SU, FSSCP and CFS feature selection algorithms. Among the averaged DRs shown in Figure 10, were achieved the highest accuracy. The average detection rate of the proposed method is 98.28, which is higher than other comparable methods. Proposed approach also has the best performance of averaged FPR shown in Figure 11. The average false positive error rate of the proposed method is 1.1, which is far lower than other methods of comparison.

8. CONCLUSION

In this paper, a hybrid filtering feature selection method is proposed for selecting the most important features in the intrusion detection standard dataset. Proposed method uses combination of information gain and symmetrical uncertainty filtering methods. Separate subsets of features from intrusion detection dataset is constructed by the application of these filtering methods and then by combining these subsets, final feature subset is generated. Proposed method overcomes the biases exist in the two filtering methods and achieves higher accuracy comparing to other filtering based methods. DR and FPR performance parameters is considered as evaluation metrics and their calculated values show that proposed method performance is higher comparing to other methods. Due to the low computational cost of filtering methods, proposed combined method obtains this performance improvement with a reasonable cost.

9. REFERENCES

[1] Gilad-Bachrach, R., Navot, A., and Tishby, N., 2004. Margin Based Feature Selection - theory and Algorithms, Proceedings of the twenty-first international conference on Machine learning (ICML), Page(s) 43–50.

[2] Wang, H., Khoshgoftaar, T.M., and Gao, k., 2010. Ensemble Feature Selection Technique for Software Quality Classification, In Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering, Page(s) 215–220.

[3] Liu, H., Li, J., and Wong, L., 2002. A Comparative Study on Feature Selection and Classification Methods Using Expression Profiles and Proteomic

Patterns, Genome Informatics, Volume 13, Page(s) 51–60.

[4] Ruiz, R., Aguilar-Ruiz, J.S., Santos, J.C., and Diaz-Diaz, N., 2005. Analysis of Feature Rankings for Classification, In Proceedings of the 6th International Symposium on Intelligent Data Analysis, Volume 3646, Page(s) 362–372.

[5] Hall M.A., and Holmes, G., 2003. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, IEEE Transactions on Knowledge and Data Engineering, Volume 15, No. 3, Page(s) 1437–1447.

[6] John, G.H., Kohavi, R., and Pfleger, K., 1994. Irrelevant Features and the Subset Selection Problem, Proceedings of the Eleventh International Conference, Page(s) 121-129, Morgan Kaufmann.

[7] Biesiada, J., and Duch, W., 2005. Feature Selection for High-Dimensional Data: A Kolmogorov-Smirnov Correlation-Based Filter Solution, in Proceedings of the 4th International Conference on Computer Recognition Systems.

[8] Witten, I.H., and Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems).

[9] Fayyad U.M., and Irani, K.B., 1992. On the Handling of Continuous-Valued Attributes in Decision Tree Generation, Machine Learning, Page(s) 87–102.

[10] Yang Y., and Pedersen, J.O., 1997. A Comparative Study on Feature Selection in Text Categorizatio, 14th International Conference on Machine Learning, Page(s) 412–420.[11] Scarfone, K., and Mell, P., 2007. Guide to Intrusion Detection and Prevention Systems (IDPS), Computer Security Resource Center (National Institute of Standards and Technology).

[11] Intrusion Detection System, http://www.webopedia.com/Term/I/intrusion_detection_system.html, [Accessed:Jan.10, 2013].

[12] Verwoerd, T., and Hunt, R., 2002. Intrusion Detection Techniques and Approaches, Computer Communications. Volume 25, Page(s) 1356-1365.

[13] Chou, T.S., Yen K.K., and Luo, J., 2007. Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms, International Journal of Computational Intelligence, Page(s) 196-208, Volume 4, No. 3.

[14] Novakovic, J., Strabac, P., and Bulatovic, D., 2011. Toward Optimal Feature Selection Using Ranking Method and Classification Alghorithms, Yugoslav Journal of Operations Research, Volume 21, No. 1, Page(s) 119-135.

[15] Hall, M.A., and Smith, L.A., 1998. Practical Feature Subset Selection for Machine Learning, Computer Science Proceedings of the 21st Australasian Computer Science Conference ACSC, Page(s) 181-191.

[16] Knowledge Discovery in Databases DARPA archive Task Description.KDDCUP 1999 DataSet, http://kdd.ics.uci. edu/databases/kddcup99/kddcup99.html, [Accessed: Dec. 06, 2012].