Hybrid feature selection for supporting lightweight intrusion detection systems

Jianglong Song, Wentao Zhao*, Qiang Liu and Xin Wang

College of Computer, National University of Defense Technology, Changsha, China
Email: songjl@nudt.edu.cn, wtzhao@nudt.edu.cn, qiangliu06@nudt.edu.cn, dywangxin@foxmail.com

Abstract. Redundant and irrelevant features not only cause high resource consumption but also degrade the performance of Intrusion Detection Systems (IDS), especially when coping with big data. These features slow down the process of training and testing in network traffic classification. Therefore, a hybrid feature selection approach in combination with wrapper and filter selection is designed in this paper to build a lightweight intrusion detection system. Two main phases are involved in this method. The first phase conducts a preliminary search for an optimal subset of features, in which the chi-square feature selection is utilized. The selected set of features from the previous phase is further refined in the second phase in a wrapper manner, in which the Random Forest (RF) is used to guide the selection process and retain an optimized set of features. After that, we build an RF-based detection model and make a fair comparison with other approaches. The experimental results on NSL-KDD datasets show that our approach results are in higher detection accuracy as well as faster training and testing processes.

1. Introduction

Intrusion detection is the process of observing the events occurred in the network or computer systems and examining them for signs of possible events that violate the security policies or security standard practices [1]. Intrusion Detection Systems (IDS) have become important and widely-used tools for ensuring network security. Since the amount of audit data that an IDS needs to examine is very large even for a small network, analysis work becomes difficult for harder detection of suspicious behaviour pattern with extraneous features [2]. IDS must reduce the amount of data to be processed to improve detection efficiency, which desires a lightweight detection system. Feature selection is one of the important and frequently-used techniques in data preprocessing for intrusion detection systems [3]. It reduces the number of features, removes irrelevant, redundant or noisy data, and brings the immediate effects for applications.

Many literatures have tried to figure out important features in order to minimize overhead of detection models and maximize detection rates. Dong et al. [4] proposed a feature selection method based on Genetic Algorithm (GA) and Support Vector Machine (SVM). Hai et al. [5] proposed an automatic feature selection procedure based on Correlation Feature Selection (CFS). Zhu et al. [6] a binary feature selection (BFS) framework in kernel spaces, where each feature is projected into kernel spaces and a binary classification task is constructed in this space. Ambusaidi et al. [7] proposed a mutual information based algorithm that analytically selects the optimal features for intrusion detection. Tama et al. [8] combined particle swarm optimization and correlation-based feature selection (PSO-CFS) with tree-based classifiers for detection task. Thaseen et al. [9] designed an intrusion detection model using fusion of chi-square feature selection and multi class SVM. Dong et al. [10] proposed a new approach to build lightweight intrusion detection system based on RF with a...
measure of feature importance. Besides this, some PCA (Principal Component Analysis) and ICA (Independent Component Analysis) approaches also have been proposed to decrease overhead of IDS and increase the detection rates [11].

Those methods can fall into two categories: filter and wrapper. Filter methods utilize the underlying characteristics of the training data to evaluate the relevance of the features by some independent measures such as distance measure, correlation measures and consistency measures [12][13]. It is more lightweight than wrapper method in terms of computation time and overheads since it is performed independent of classification algorithms. However, their results are not always acceptable. Due to the lack of interaction between the classifier and the dependence among features, filter methods might fail to choose the best available subset or might select redundant features [14]. Wrapper method exploits a machine learning algorithm as a fitness function to evaluate the goodness of features. It provides better performance of selecting suitable features since it employs performance of learning algorithm as an evaluation criterion. In comparison with filter methods, wrapper methods are argued to be more accurate. However, the wrapper approaches are often computationally more complicated when dealing with large sets of features than the filter approaches [15].

To cope with the aforementioned drawbacks, we propose to adopt the principle of chi-square feature selection and RF algorithm to build a lightweight intrusion detection model in a hybrid manner. The proposed approach is able not only to significantly decrease training and testing time while retaining high detection rates but also to figure out important features simultaneously. Then, we design an RF-based detection model framework with the proposed hybrid feature selection method and conduct several experiments on intrusion detection benchmark. The experimental results on NSL-KDD datasets indicate the feasibility of our method.

This paper is organized as follows. In Section 2, the proposed hybrid feature selection approach is introduced. Experiments and results are presented in Section 3. Finally, the paper is concluded in Section 4.

2. Hybrid Feature Selection Method

In this section, we introduce the proposed hybrid feature selection method which utilizes both filter and wrapper methods. The framework of the approach is depicted in figure 1. It consists two stages: the first stage at which the chi-square statistics is used for feature ranking and elimination, and the second stage which determines the optimal subset using RF algorithm, and contributes maximum classification accuracy on training dataset.

2.1. Filter Method for Feature Pre-Selection

The $\chi^2$ method [16] evaluates features individually by measuring their chi-square statistic with respect to the class labels. The $\chi^2$ value of a feature is defined as:

$$\chi^2 = \sum_{i} \sum_{j} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

(1)

where

$$E_{ij} = R_i \times L_j / S$$

(2)

$t$ is the number of different values in a feature, $l$ is the number of class labels, $A_{ij}$ is the number of samples in the $i$-th feature value and $j$-th class, $R_i$ is the number of samples in the $i$-th feature value, $L_j$ is the number of samples in the $j$-th class, $S$ is the total number of samples and $E_{ij}$ is the expected frequency of $A_{ij}$. 

After calculating the $\chi^2$ value of all considered features in each sample, these values can be sorted with the largest one at the first position, as the larger the $\chi^2$ value, the more important the feature is. This will provide an ordered ranking of features, and a threshold needs to be established. The filtering process is applied to eliminate irrelevant and redundant features from the initial set. This helps the wrapper method-based feature selection to narrow the searching range from the entire original feature space to the pre-selected features.

2.2. Wrapper Feature Selection Using RF
Random Forests builds an ensemble of CART (Classification and Regression Tree) tree classifications using bagging mechanism [17]. By using bagging, each node of trees only selects a small subset of features for the split, which enables the algorithm to create classifiers for high dimensional data very quickly. This somewhat counterintuitive strategy turns out to perform very well, compared with the state-of-the-art methods in classification and regression. Also, RF produces additional facilities, especially the feature importance by numerical values [18].

Once the filter method finishes its task, the second stage evaluates the candidate feature subsets, using wrapper scheme, to determine the optimal subset of feature that can produce the best classification performance. To do so, RF and the classification accuracy are employed. If the performance reaches the best accuracy rate, the selection process is completed while outputs the last optimal subset of features. The procedure is specified as follows:

(i) Delete one input attribute from the data.
(ii) The resultant data are used for training.
(iii) The results of the classifier are analyzed using the performance metrics.
(iv) The rules are used to rank the attribute by its importance level.
(v) Repeat the steps (i) to (IV) for each of the features.

2.3. RF-Based Intrusion Detection Framework
In this paper, for demonstrating the efficiency and effectiveness of our approach, we build an intrusion detection model based on RF using the features selected by our proposed hybrid feature selection method. The overall architecture of the model is depicted in figure 2.
3. Experiments and Results
All experiments were performed in a Win7 OS having configurations Intel(R) Core(TM) i7-4790K CPU@4.00GZ, 16GB RAM. We have used RF version in an open machine learning library scikit-learn using Python programming language [19] [20].

3.1. Description of the Benchmark Dataset
The NSL-KDD dataset was proposed by Tavallaee et al. [21], which is a new revised version of the KDD Cup 99 dataset. The important deficiency in the KDD Cup 99 dataset is the huge number of redundant records that will cause learning algorithms to be biased towards the more frequent records, and thus prevent it from learning infrequent records. The existence of these repeated records in the test set, on the other hand, will cause the evaluation results to be biased by the methods which have better detection rates on the frequent records [21]. The provided NSL-KDD dataset does not suffer from any of the mentioned problems. Furthermore, the number of records in the train and test sets is reasonable. This advantage makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research work will be consistent and comparable.

The total number of records in the training dataset is 125973, where 67343 records are normal data and the rest indicate attacks. The total number of features is 41, which include numerical, nominal, and binary features. This dataset consists of five different classes, where one shows normal behavior and the rest indicate attacks. Attacks are categorized as DoS, Probe, R2L, and U2R. The test set consists of 22544 records. In this paper, 20% percent of the full training data is applied to the proposed method. Furthermore, to validate the performance of our hybrid feature selection, we test our RF-based detection model with the proposed hybrid feature selection on KDDTest-21 dataset and make a comparison with other detection model.

3.2. Performance Metrics
In this paper, we classify the dataset into two main classes: normal and abnormal. In order to classify them, we use RF to train a binary classifier. Several experiments have been conducted to evaluate the performance of the proposed method. For this purpose, the accuracy rate, detection rate, and F-measure metric are applied, and they are defined by:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}
\]

(3)

\[
\text{Detection Rate} = \frac{TP}{TP + FN}
\]

(4)

where True Positive (TP) is the number of actual attacks classified as attacks, True Negative (TN) is the number of actual normal records classified as normal ones, False Positive (FP) is the number of actual normal records classified as attacks, and False Negative (FN) is the number of actual attacks classified as normal.

The F-measure metric is a statistical technique for examining the accuracy of a system by considering both precision and recall of the system. F-measure used in this paper assigns the same
weights to both Precision and Recall, and is given by:

$$F - \text{measure} = \frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

(5)

The Precision is the ratio of items correctly classified as X to all items classified as X. A higher value of precision means a lower false positive rate and vice versa. The Recall is another important value for measuring the performance of the detection system and to indicate ratio of items correctly classified as X to all items that belong to class X [22].

$$\text{Precision} = \frac{TP}{TP + FP}$$

(6)

$$\text{Recall} = \frac{TP}{TP + FN}$$

(7)

3.3. Experimental Results and Analysis

We compare the results of the RF-based detection model using the proposed hybrid feature selection method with the detection model using all 41 features and the chi-square filter algorithm only. Features selected by the aforementioned three feature selection methods are listed in Table 1. Table 2 summarizes the classification results of the different selection methods with respect to detection rate, accuracy and F-measure. Through Table 2, we can see that our detection model with the proposed hybrid method achieve the highest accuracy rates of 80.07%. In addition, the proposed approach achieve detection rate of 71.21%. In general, in terms of the F-measure results for all methods, the detection model with the proposed hybrid feature selection enjoys higher rates.

Table 3 shows the average training and testing time of the detection model with the proposed hybrid feature selection, compared with using only filter method and those using all 41 features. It can be observed that the proposed approach illustrates the best average time of building and testing processes.

### Table 1. Feature importance with three different feature selection methods.

| Algorithm           | #  | Feature ranking                  |
|---------------------|----|----------------------------------|
| Proposed hybrid method | 7  | f5, f11, f4, f3, f2, f29, f35    |
| Filter method       | 30 | f4, f5, f33, f11, f3, f1, f34, f38, f2, f31, f22, f35, f27, f32, f29... |
| All features        | 41 | f3, f4, f28, f5, f33, f29, f25, f2, f36, f31, f1, f24, f22, f11, f32, f35, f23... |

### Table 2. Performance of classification on the KDD Test+ dataset.

| RF model with: | Detection rate | Accuracy | F-measure |
|---------------|---------------|----------|-----------|
| Proposed hybrid method | 71.21 | 80.07 | 80.27 |
| Filter method | 64.49 | 78.57 | 77.41 |
| All features | 65.71 | 79.23 | 78.28 |

### Table 3. Average time of building and testing processes on the KDD Test+ dataset.

| RF model with: | Accuracy time |
|---------------|---------------|
| Building time(ms) |               |
| Proposed hybrid method | 110 |
| Filter method | 156 |
| All features | 172 |
| Testing time(ms) |           |
| Proposed hybrid method | 160 |
| Filter method | 310 |
| All features | 330 |

In order to demonstrate the performance of the RF-based model with our hybrid feature selection,
experiments have been conducted to make comparisons with some state-of-the-art detection model on the KDD Test+ dataset. The results illustrate in figure 3 strongly indicate that the proposed detection model shows promising results compared with other models. Regarding the results obtained by [21], it can be seen that the proposed approach enjoys the best accuracy and our proposed method can significantly improve the classification performance (up to 12.38 % improved) compared with other six competitors.

![Figure 3. The detection accuracy of different models on the KDDTest-21 dataset](image)

4. Conclusion
In this paper, for building a lightweight intrusion detection model, a hybrid feature selection approach combining the filter and wrapper selection processes is proposed to achieve more accurate detection as well as fast training and testing process. It contains two stages: (1) chi-square filter feature eliminating phase; and (2) wrapper feature selection using RF. Experiments on the NSL-KDD Test+ dataset exhibit promising results in terms of classification accuracy, low computational cost and F-measure. In addition, compared with those models that have been evaluated on the NSL-KDD Test-21 dataset, the RF-based detection model with the proposed hybrid method shows good a comparable result in terms of detection accuracy.

5. Acknowledgments
This work is partly sponsored by the National Defense Key Laboratory Fund Project under Grant No.1602068614001.

6. References
[1] Barry B I A and Chan H A 2011 Intrusion Detection Systems (Berlin, Heidelberg: Springer)
[2] Lee W, Stolfo S J and Mok K W 1999 Proc. of the 1999 IEEE Symp. on Security and Privacy (Oakland) vol 00 (Los Alamitos: IEEE Computer Society) pp 120–132
[3] Chen Y, Li Y, Cheng X Q and Guo L 2006 Current Medical Research and Opinion 44 1309–18
[4] Dong S K, Nguyen H N, Ohn S Y and Park J S 2005 Advances in Neural Networks - ISNN 2005, 2nd Int. Symp. on Neural Networks (Chongqing) vol 3498 (Berlin, Heidelberg: Springer) pp 415–420
[5] Nguyen H T, Franke K and Petrovic S 2011 Int. J. Mob. Comput. Multimed. Commun. 3 21–34
[6] Zhu C, Liu X, Zhou S and Liu Q 2014 *Int. Joint Conf. on Neural Networks* (Beijing: IEEE) pp 4190–97

[7] Ambusaidi M A, He X, Nanda P and Tan Z 2016 *IEEE Transactions on Computers* **65** 2986–98

[8] Tama B A and Rhee K H 2015 *Advances in Computer Science and Ubiquitous Computing* vol 373 ed Park D S, Chao H C and et al (Singapore: Springer) pp 489–495

[9] Thaseen I S and Kumar C A 2016 *Journal of King Saud University - Computer and Information Sciences* (2016) 003

[10] Dong S K, Sang M L and Park J S 2006 *Lecture Notes in Computer Science* **3973** 224–230

[11] Cataltepe Z, Genc H and Pearson T 2007 *15th European Signal Processing Conf.* (Poznan: IEEE) pp 970–974

[12] Dash M, Liu H and Motoda H 2000 *4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, Current Issues and New Applications* vol 1805 ed Terano T, Liu H and et al (Berlin, Heidelberg: Springer) pp 98–109

[13] Almuallim H and Dietterich T G 1994 *Artificial Intelligence* **69** 279–305

[14] Peng Y, Wu Z and Jiang J 2010 *Journal of Biomedical Informatics* **43** 15–23

[15] Huang J, Cai Y and Xu X 2006 *18th Int. Conf. on Pattern Recognition* vol 2 (Hong Kong: IEEE) pp 618–621

[16] Liu H and Setiono R 1995 *Proc. of 7th IEEE Int. Conf. on Tools with Artificial Intelligence* vol 00 (Los Alamitos: IEEE Computer Society) pp 388–391

[17] Duda R O, Hart P E and Stork D G 2000 *Pattern Classification (2nd Edition)* (New York: Wiley-Interscience)

[18] Cutler A, Cutler D R and Stevens J R 2012 *Ensemble Machine Learning: Methods and Applications* ed Zhang C and Ma Y (Boston: Springer) pp 157–175

[19] Pedregosa F and et al 2011 *Journal of Machine Learning Research* **12** 2825–30

[20] Buitinck L and et al 2013 *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* vol 238 pp 108–122

[21] Tavallaee M, Bagheri E, Lu W and Ghorbani A A 2009 *Proc. of the 2nd IEEE Int. Conf. on Computational Intelligence for Security and Defense Applications* (New York: IEEE) pp 53–58

[22] Buczak A L and Guven E 2016 *IEEE Communications Surveys Tutorials* **18** 1153–76