OPEN ACCESS 32

ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 5, Issue. 2, December 2019. pp. 32-39
doi: 10.21108/IJOICT.2019.52.209

# Wrapper-Based Feature Selection Analysis For Semi-Supervised Anomaly Based Intrusion Detection System

Andreas Jonathan Silaban[1], Satria Mandala[1], Erwid Jadied Mustofa[1*]

[1] *School of Computing, Telkom University*
*Bandung, Indonesia*

[*] jadied@telkomuniversity.ac.id

**Abstract**

Intrusion Detection System (IDS) plays as a role in detecting various types of attacks on computer networks. IDS identifies attacks based on a classification data network. The result of accuracy was weak in past research. To solve this problem, this research proposes using a wrapper feature selection method to improve accuracy detection. Wrapper-Feature selection works in the preprocessing stage to eliminate features. Then it will be clustering using a semi-supervised method. The semi-supervised method divided into two steps. There are supervised random forest and unsupervised using Kmeans. The results of each supervised and unsupervised will be ensembling using linear and logistic regression. The combination of wrapper and semi-supervised will get the maximum result.

**Keywords**: intrusion detection system, semi-supervised, wrapper, feature selection

## I. INTRODUCTION

NETWORK security becomes the most important and critical service in the information technology era. Intrusion Detection System (IDS) plays as a role in detecting various types of attacks in computer networks [Sharma T, 2011]. Since the attacks get more complex and complicated, security needs to improve the process. This research proposed Semi-Supervised Anomaly-based IDS. It works by executing the label and unlabeled data and classify them into each type of attack. To optimize the result, semi-supervised will combine with wrapper feature selection. Wrappers can increase accuracy by eliminating the useless feature in the dataset.

The wrapper method evaluates the subset of variables that possible to use. Two main disadvantages of this method have increased the risk of overfitting and take too much time for large data. Overfitting is a small error ratio when ruins in training data. This causes a problem because it's not realistic to predict unknown data [T. M. Phuong, 2005]. To solve this problem, the existing wrapper method has been applied to some research [S. Revathi, 2013] [B. Senthilnayaki, 2015].

The feature selection from the existing wrapper method isn't optimal to decrease time execution and improves accuracy. An efficient preprocessing phase and a suitable optimization algorithm will optimize the classification of attacks.

Received on Feb 2018. Accepted on Jun 2020

This research focuses on the wrapper feature selection, which aims to learn how wrapper algorithms work, how much the optimization algorithm impact on accuracy. The wrapper is the best.

## II. LITERATURE REVIEW

The Research of Feature Selection on Intrusion Detection System for almost ten years showed many wrapper methods implemented in IDS. However, based on the last research still has an accuracy problem. The accuracy wasn't optimal. This research will solve those problems by using the wrapper and semi-supervised working together.

In early 2003, Alexander Hofinann [Alexander Hofinann, 2004] proposed the evolution of the wrapper method. He claims this wrapper is an evolutionary algorithm feature selection. However, when the wrapper implemented in Naive Bayes' classification, the result was still low around 75 percent. He implements the wrapper in the J48 classification, and the result was around 87 percent. A year later, in 2004, Shital C. Shah et al. [Shah, 2004] researched the genetic algorithm based on gene selection. The authors proposed the wrapper method with the Genetic Algorithm search algorithm and decision tree classification. The author optimized the wrapper method by modifying the GA search algorithm. But, the result of accuracy is 55.06 percent. This result is getting worse than the last research. From the research above, GA doesn't fit with the wrapper method. Five years later, in 2009, Li-Yeh Chuang et al. [Chuang, 2009] researched about Tabu Search and Binary Particle Swarm Optimization for Feature Selection Using Microarray Data. The authors proposed a wrapper method that used a search + PSO search algorithm with KNN and SVM classification. While evaluation function using euclidean distance. The results obtained on the search taboo method with KNN classification were 84.8 percent. While the results obtained on the TS + PSO method with a KNN classification of 91.19 percent and for SVM 94.30 percent, it's good enough than before. In 2014, Yudong Zhang et al. [Uguz, 2014] curious enough with the last research about PSO and wrapper. So, the authors researched Binary PSO with a mutation operator for feature selection using a decision tree to spam detection. The author proposed a wrapper method using binary search algorithm and decision tree classification. The author explained the wrapper has some weaknesses. There are slow execution and lack of generality. This problem can be solved with global optimization and using K- fold cross-validation. The new PSO algorithm is capable of increasing accuracy up to 93.34 percent compared to the GA 92.07 percent algorithm. In the same year, Mohammed A. Ambusaidi et al. [Ambusaidi, 2014] researched about feature selection for intrusion detection data classification by using filter and wrapper. The first step is the filter method performing feature ranking and eliminating the irrelevant feature that aims to reduce the cost of wrapper computing. Then the second step is the IFFS-based wrapper method aims to find a subset that can improve accuracy. The final subset is classified by using the SVM technique with the KDD'99 dataset. The author claims this method can improve accuracy on IDS. The accuracy rate achieved is 99.1 percent.

TABLE I
LIST OF EARLIER RESEARCH

| Writer | Dataset | Algorithm | Classification | Accuracy |
|---|---|---|---|---|
| S. Revathi, Dr. A. Malathi | NSL-KDD | CFS | *Naïve Bayes, J48* | 75%, 87% |
| Shital C. Shah, Andrew Kusiak | *Drug, Placebo* | *Genetic Algorithm* | *Decision Tree* | 51.33%, 55.06% |
| Yudong Zhang, Shuihua Wanga, Preetha Phillips, Genlin Ji | Spam | Binary PSO | *Decision Tree* | 92.07% |

| | | | | |
|---|---|---|---|---|
| Mohammed A. Ambusaidi, Xiangjian He, Zhiyuan Tan, Priyadarsi Nanda, Liang Fu Lu and Upasana T.Nagar | KDD'99 | IFFS | SVM | 99.1% |
| Li-Yeh Chuang, Cheng-Huei Yang, Cheng-Hong Yang | *Microarray* | *Tabu Search + PSO* | KNN, SVM | 91.19%, 94.3% |

An intrusion detection system (IDS) can be signature-based or anomaly-based. This research focused on anomaly-based IDS to improve its performance

### *A. Intrusion Detection System*

IDS handles many varieties of network traffic patterns that contained large data. Each pattern on the dataset is a unified feature [L. Dhanabal, 2006]. IDS is responsible for classifying attacks based on network activity data. There are three stages in IDS:

1. *Preprocessing*
   In this phase, the input dataset is converted into a numeric (binary) form. Then performed the feature selection process to improve the dimensionality of the dataset.
2. *Classification*
   After the preprocessing stage, the attacks classified in each category of types.
3. *Output*
   Anomaly detection can detect unrecognized attacks, but it caused the increase of False Positive Rate (FPR). It because the network traffic anomaly can occur due to an error or technical error, but IDS detects it as an attack or suspicious activity.

### *B. Machine Learning*

Machine learning is the science that learns about design and development algorithms that allow computers to develop behaviors based on empirical data, such as database data sensors. Based on the input and output of the algorithm, the types of an algorithm in machine learning can be grouped into:

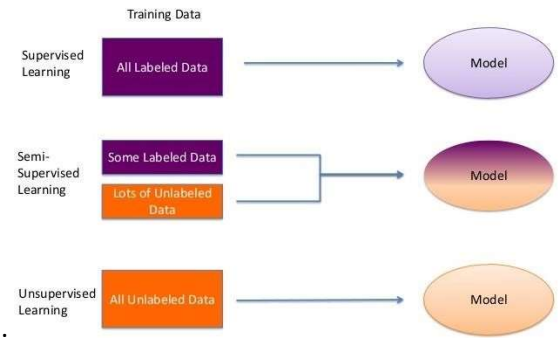


Fig 1. Clustering

1. *Supervised Learning* is the machine learning task of inferring a function from labeled training data. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.
2. *Unsupervised Learning* is the machine learning task of inferring a function to describe hidden structure from unlabeled data.
3. *Semi-Supervised Learning* is a class of supervised learning tasks and techniques that also make use of unlabeled data for training. Typically a small amount of labeled data with a large amount of unlabeled data.

### C. *Feature Selection*

Feature Selection is the process of selecting a feature subset from many features to reduce the dimensionality of the dataset [Alexander Hofinann, 2004]. Feature Selection is an effective method for IDS classification performance. Subset feature reductions reduce the complexity of the time and improve the accuracy of the attacker's denominator. Based on the method of implementation, feature selection is divided into a rank selection and subset selection.

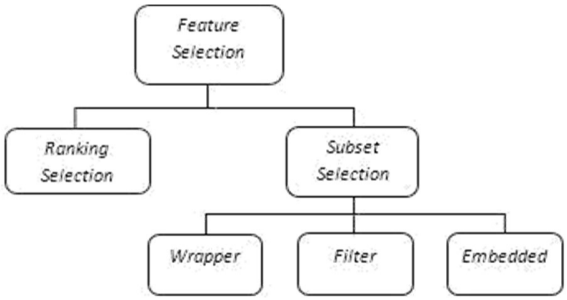


Fig 2. Feature Selection

1. *Wrapper* method process the selection of features along with modeling execution. This type of selection uses a criterion that utilizes the classification rate of the classification or modeling method used. To reduce computational cost, the selection process is generally done by utilizing the classification rate of the method of classification or modeling classification for modeling with the lowest value. For the wrapper type, it is necessary to perform the feature subset selection before determining which subset is the best-ranked subset.
2. *Filter* method is similar to wrapper selection by using intrinsic statistical properties of the data. The filter type differs from the wrapper type in the feature review, which is not done simultaneously with the modeling performed. Selection is performed by utilizing one of several types of filters that exist. The method of selecting this filter is generally performed at the preprocessing stage and has a low computational cost.
3. *Embedded* utilizes a learning machine in the feature selection process. In this selection system, the feature is naturally eliminated, if the learning machine considers the feature not so influential. Some learning machines that can be used include Decision Trees, Random Forests, and others.

## III. RESEARCH METHOD

An intrusion detection system (IDS) can be signature-based or anomaly-based. This research focused on anomaly-based IDS to improve its performance. There are many steps to this plot:

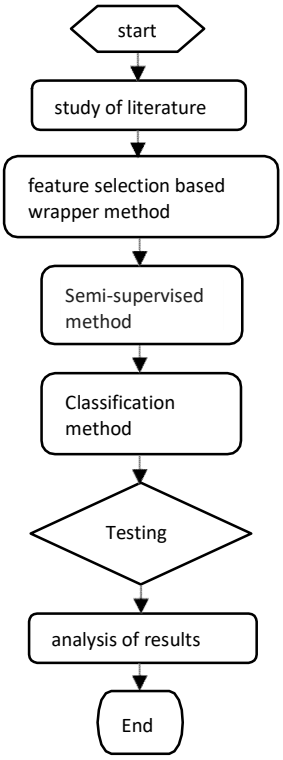


Fig 3. Flowcart Diagram

1. Study of literature, at this stage, the authors learn about literature studies, the authors studied the existing wrapper and then analyzed how to optimize it. The literature study was conducted by reading 15 papers on feature selection, IDS, and datasets.
2. Feature Selection-based Wrapper, at this step, the authors used a wrapper to execute some features in the dataset.
3. The semi-supervised method, at this step, the dataset learned into supervised and unsupervised data.
4. Classification using Random Forest when clustering data.
5. Testing If the test result is better than the existing method, then it will be continued to the analysis phase of the result. However, if the performance of the proposed method doesn't reach the objective, then the experiments will be re-done to the NSL-KDD dataset with modification or algorithm, which will then be tested for performance through testing based on accuracy matrix and time complexity.
6. Analysis of Results When it meets the proposed matrix, the new method is analyzed to find the dataset feature selection pattern applied to the preprocessing phase.

### *1. Data*

The data studied and identified in the proposed experiments is the NSL-KDD dataset. The NSL-KDD dataset is the most commonly used data source as an object of study in IDS development [Mrutyunjaya, 2007]. NSL-KDD contains an anomaly detection data record. The NSL-KDD dataset includes 41 features, five normal classes, and four attack types (Dos, Probe, R2L, and U2R). The NSL-KDD dataset is the latest version of the KDD'99 dataset that has been filtered from redundant data and duplicated data.

Although the number of data is less, the data processing time is still not efficient because of the amount of data structures that must be processed [L.Dhanabal, 2006]. The structure of the NSL-KDD dataset has various data attributes. The completeness of the processed data attribute does not guarantee accurate processing and requires a long processing time. The data processing time is due to the inefficient attribute structure.

TABLE I
LOAD DATA

| *Data* | *Count* |
|---|---|
| Normal | 67343 |
| DoS | 45927 |
| Probe | 11656 |
| R2L | 995 |
| U2R | 52 |

*2. Metric*

The metric that will be the performance comparison points between the proposed method and the comparison method is the accuracy of classification (AC) This metric is a benchmark on the accuracy results of existing methods with newly developed methods. To calculate the accuracy, the following formula is used.

$$\frac{TP\ +\ TN}{TP\ +\ TN\ +\ FP\ +\ FN} = \frac{TP + TN}{N} \tag{1}$$

$$\frac{FP}{N} = \frac{FP}{FP + TN} \tag{2}$$

Description :

- TP : True Positive
- FP : False Positive
- TN : True Negative
- FN : False Negative
- N : Total Accuracy

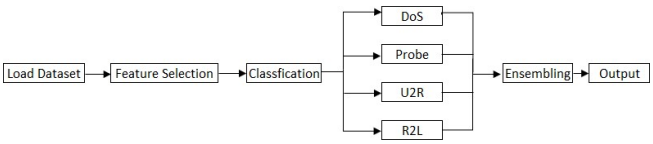


Fig 4. Flowchart Diagram

- Dataset, at this stage, the dataset used is part of the NSL-KDD. NSL-KDD has 41 features, five classes, and four types of attack. NSL-KDD is punctuated by equalizing the data format to be processed by the wrapper.
- Feature Selection Wrapper Method, 41 features in NSL-KDD will be selected in this process. The feature selection process is done by the wrapper method. The selected feature is a feature that is rated as having maximum contribution to the test.

- Classification, in this stage, the feature will be categorized using techniques JST and KNN. The selected feature will be applied to the four types of attacks available on the NSL-KDD. The feature guide is run on DOS, Probe, U2R, and R2L attacks.
- Evaluation, in this stage, the test will be obtained the results of accuracy, and a long time-ward is required. The results of accuracy and detection time obtained are reviewed against preliminary data. After doing the review, then examined how much increase happened.
- Output, in this stage, the final results obtained to become the comparative value of the results of expert research, existing methods, and methods developed. To get concrete results then the test will be done several times

## IV. RESULTS AND DISCUSSION

The test results were done using python with the classification of Random Forest. Based on the results of tests conducted, the value of Kmeans is better than Gaussian Mixture when clustering the unsupervised data. The result of unsupervised test ensembling with the supervised result. Below is the table of test results

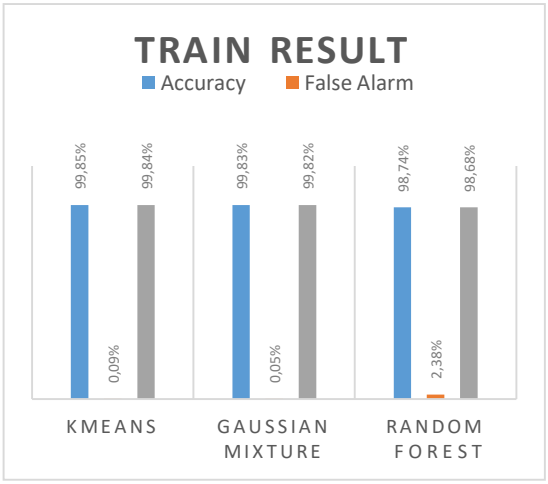


Fig 5. Train Result

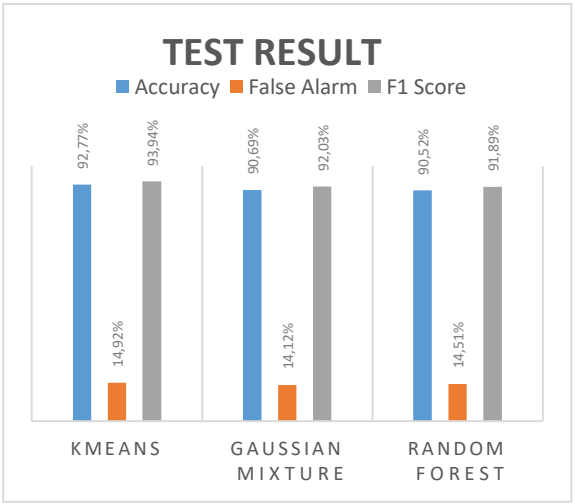


Fig 6. Test Result

Now, we have the result from the data test. From the graph above, the best single method for unsupervised clustering is Kmeans. It ensembles with supervised clustering Random Forest. To get the final result, we use a linear combination and Logistic Regression with Random Forest Classifier.

TABLE III
ENSEMBLING METHOD

| | *Accuracy* | *False Alarm* | *F1 Score* |
|---|---|---|---|
| Linear Combination | 92.645% | 16.1672% | 93.8927% |
| Logistic Regression | 92.618% | 16.6512% | 93.8904% |
| Stacking Random Forest | 92.641% | 16.1775% | 93.8893% |

## V. Conclusion

From the results of the research conducted, the best single method is Kmeans when compared with Gaussian Mixture. This results in unsupervised data. To obtain labeled and unlabeled data hail, it will be combined with a supervised random forest. To combine it using ensembling technique with linear combination stages, then logistic regression and last stacking with random forest.

## ACKNOWLEDGMENT

The authors would thank our institution Telkom University who provided facilities for this research and support us doing this research. We would also like to show our gratitude to references for sharing their research so we can completely finish this research, and we thank our colleague graham and Reza, who help us while doing this research. Finally, with all of our respect, we hope this research can have a real impact in the real world.

## REFERENCES

1. Sharma, T., and Sinha, K. Intrusion detection system Technology, 2011.
2. T. M. Phuong, Z. Lin et R. B. Altman. Choosing SNPs using feature selection. Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference, pages 301-309, 2005. PMID 16447987
3. S. Revathi, D. A. M. A detailed analysis on nsl-kdd dataset using various machine learning techniques for intrusion detection. International Journal of Engineering Research Technology (IJERT) 2 (2013), 1-3.
4. B. Senthilnayaki, Dr.K.Venkatalakshmi, Dr.A.Kannan, Intrusion Detection Using Optimal Genetic Feature Selection and SVM based Classifier, 2015.
5. Alexander Hofinann." Feature Selection for Intrusion Detection: An Evolutionary Wrapper Approach". IEEE transactions on systems applications, 2004.
6. Mrutyunjaya Panda, M. R. P. Network intrusion detection using naIve bayes. IJCSNS International Journal of Computer Science and Network Security, 1-6.
7. Shah, S. C.; Kusiak, A. (2004). "Data mining and genetic algorithm based gene/SNP selection". Artificial intelligence in medicine. 31 (3): 183–196
8. Chuang, L.-Y.; Yang, C.-H "Tabu search and binary particle swarm optimization for feature selection using microarray data". Journal of computational biology. 16(12): 1689–1703, 2009
9. Uguz, H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowl.-Based Syst 24 (2011), 1024-1032
10. L.Dhanabal, D. S. S. A study on nsl-kdd dataset for Intrusion detection system based on classification algorithms.1-3.