OPEN ACCESS
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Intl. Journal on ICT
Vol. 4, Issue. 2, December 2018. pp. 43-50
doi:10.21108/indojc.2018.42.216

Increasing Feature Selection Accuracy through Recursive Method in Intrusion Detection System

Andreas Jonathan [#1], Satria Mandala [#2]

School of Computing, Telkom University
Jl. Telekomunikasi No.1 Dayeuh Kolot Terusan Buah Batu Bandung
[1]anjonathan04@gmail.com
[2]satriamandala@telkomuniversity.ac.id

Abstract

Artificial intelligence semi supervised-based network intrusion detection system detects and identifies various types of attacks on network data using several steps, such as: data preprocessing, feature extraction, and classification. In this detection, the feature extraction is used for identifying features of attacks from the data; meanwhile the classification is applied for determining the type of attacks. Increasing the network data directly causes slow response time and low accuracy of the IDS. This research studies the implementation of wrapped-based and several classification algorithms to shorten the time of detection and increase accuracy. The wrapper is expected to select the best features of attacks in order to shorten the detection time while increasing the accuracy of detection. In line with this goal, this research also studies the effect of parameters used in the classification algorithms of the IDS. The experiment results show that wrapper is 81.275%. The result is higher than the method without wrapping which is 46.027%..

Keywords: intrusion detection system, recursive method, wrapper, feature selection, accuracy

I. Introduction

NETWORK securities becomes the most important and critical services in information technology era. Intrusion Detection System (IDS) plays as a role in detecting various types of attacks in computer networks [Sharma T, 2011]. However, The weakness of IDS is the accuracy is not optimal. This problem will be the focus of this research. Based on the previous research, the caused of long processing time is due to the large number of features. To reduce the feature then this research proposes wrapper feature selection. Wrapper works by reduces existing features and sorted selected features by rank. Reduced features will be process in the semi-supervised learning. Semi-supervised IDS can handle a variety of attacks both label and unlabel attack data.

However, the wrapper method has two main drawbacks, namely the increased risk of overfitting and long processing times for large data. Overfitting is the model error ratio is very small when compared with training data [2]. The weakness of the wrapper was proven in this study. It takes longer to process the semi-supervised wrapper IDS than without using a wrapper. However, the accuracy of the wrapper is better than without a wrapper.

Received on Nov 2018. Accepted on Dec 2018

II. Literature Review

The research on feature selection during the 2003-2014 has shown that many wrapper methods have been developed. However, based on the latest research the accuracy is still not optimal. The main problem of previous research is the long processing time and the results of accuracy that are not optimal. In this study will analyze how the wrapper works to optimize the accuracy of attack detection. Wrapper research was conducted in early 2003, researcher Alexander Ho nann [5] proposed the evolution of the wrapper method. The author claims that the wrapper is an evolutionary feature selection algorithm. However, when the wrapper is implemented in the Naive Bayes classification, the results are relatively low at around 75 percent. The author tried to re-apply the wrapper in the J48 classification and the results are around 87 percent [6]. A year later Shital C. Shah et al. [7] proposed a wrapper method with a genetic algorithm search algorithm and classification of decision trees. The author optimizes the previous wrapper method by modifying the GA search algorithm. However, the accuracy has decreased by 55.06 percent. Five years later, Li-Yeh Chuang et al. [8] examined Tabu Optimization for Search and Binary Particle Swarm for Feature Selection using Microarray Data. The author applies the wrapper method with KNN and SVM classifications. Accuracy results are 91.19 percent and 94.30 percent. And then in 2014 Yudong Zhang et al. [9] continued the latest research on PSO and wrapper. The author proposes a wrapper method using binary search algorithms and classification of decision trees. In this study, the authors explained the wrapper has several weaknesses, namely the long execution time and the absence of generality. This problem can be overcome by global optimization and using K-fold cross validation. The new PSO algorithm is able to increase accuracy up to 93.34 percent compared to GA 92.07 percent algorithm. Mohammed A. Ambusaidi et al [10] continued the study using filter and wrapper methods. The final subset is classified using SVM technique with KDD99 dataset. The author claims this method can improve accuracy on IDS. The accuracy level 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%

A. Intrusion Detection System

Intrusion Detection System (IDS) is a system that control network traffic and suspicious activity within a network system. If any suspicious activity detected and related to network traffic then IDS will alert the system or network administrator. Intrusion Detection System is responsible for identify, classify, and responding to suspicious activities. IDS classified attacks based on network activity data.

B. Wrapper Feature Selection

The wrapper method reduces the feature by measuring the feature's functionality based on the performance classification. The wrapper method consists of several submethod, there are recursive feature elimination (RFE), sequential feature selection (SFS), and genetic algorithms (GA). In this research, the author using RFE. The author using RFE because it is computationally more efficient by using feature weight coefficients (linear models) or feature usefulness (tree-based algorithms) to recursively recycle features, while SFS eliminates (or adds) features based on user-specified classifier / regression performance metrics. The RFE method is a linear model that reduces features recursively by considering a smaller feature set and then sorting it by rank. RFE has a random_state parameter which is a random number generator pseudo used during data processing. The default parameter is 0 (zero). First, the estimator is trained in the initial feature set and analyzes the importance of each feature gained through the coef_ or feature_importances attribute. Then the most important feature will be selected from a set of existing features. This procedure is recursively repeated until the number of features desired to be selected is finally reached. Once achieved, the selected features will be applied to the data train and test data for testing.

C. Dataset

The dataset used in this study is NSL-KDD. NSL-KDD consists of 22 types of attacks, 4 attack categories and 41 features. The type of NSL-KDD used in the data train is KDDTrain around 20Percent, while for the test data is KDDTest +. The data train and test data will be processed at the preprocessing stage to convert the whole string to binary. The processed data will be processed at a later stage.

D. Semi Supervised

Semi-Supervised Learning is a method that combines supervised and unsupervised to produce a function. Semi supervised can process labeled or unlabelled data, making it flexible to detect existing attacks on databases or new attacks. The semi-supervised learning algorithm processes the data labeled in small amounts and unlabeled data in large quantities. Semi supervised used in this research is LabelPropagation library.

E. Support Vector Machine (SVM)

One of the classifications used is the Support Vector Machine (SVM). The SVM model is an example representation as a point in space, mapped so that examples of separate categories are divided by the clearest gaps that are as wide as possible. New samples are then mapped into the same space and predicted to fall into the category by the side of the slot in place. Since the data in this study are non-linear, kernels are required to map the feature vector into a high-dimensional space, so the nonlinearity problem can be solved linearly. The kernel to be used is Radial Basis Function (RBF). The default setting for the classification parameter is kernel = rbf and gamma= 20. To improve accuracy detection, it will perform parameter tuning in each kernel.

F. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) classification is a method to classify objects based on learning data closest to the object. Learning data is projected into many-dimensional spaces, where each dimension represents a feature of the data. This space is divided into sections based on the classification of learning data. A point in this space is marked by class c if class c is the most common classification of the nearest neighbor of the titk. Close or away neighbors are usually calculated based on Euclidean distance. Parameters on KNN are n_neighbor = 7. To improve accuracy detection, parameter tuning will be performed

III. Research Method

A. Research Framework

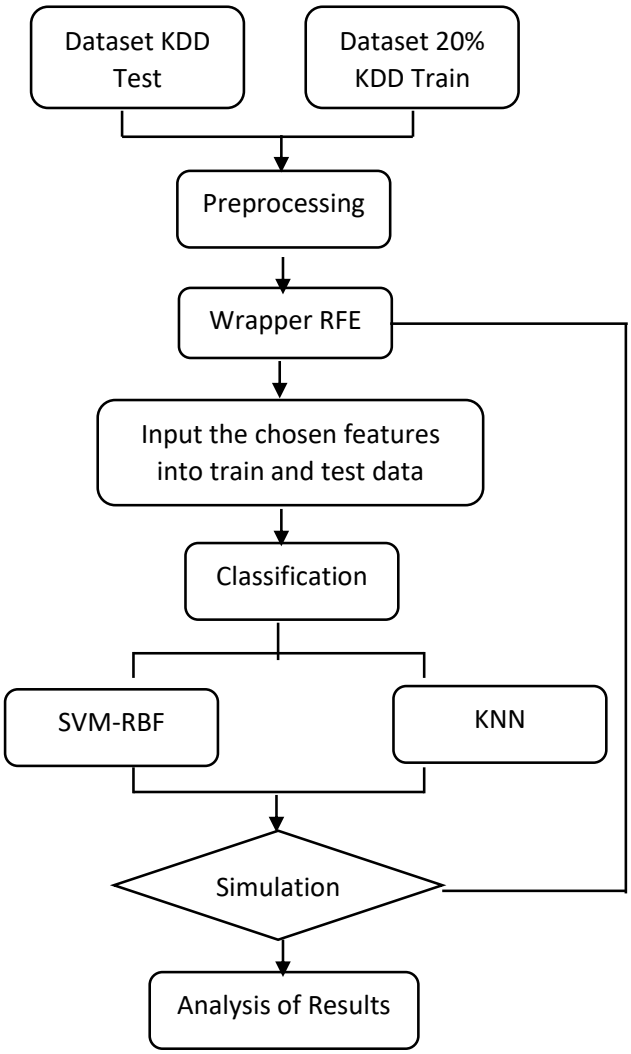

Fig 1. Flowcart Diagram

This research methodology begins at the stage of determining the dataset. The dataset used for training is KDDTrain_20Percent and the dataset used for the test is KDDTest +. The dataset will be processed at the preprocessing stage using matlab. In the preprocessing stage, all data of type string will be converted to binary, so all data contents on the dataset are binary. This processed data has a .csv output. After the data completes the preprocessing stage, then the new data is trained on the wrapper. At this stage, the wrapper learns on every feature on the datatrain. Then the wrapper reduces one by one feature according to the number of desired

features. The selected feature is then re-inserted into the datatrain and datatest as new data. The new datatrain and datatest are ready to be processed at the classification stage. At the classification stage, parameter tuning will be performed to compare which results are better. Tuning parameters are performed on SVM-RBF with gamma parameter, and KNN with parameter n_neighbors. After entering the value you want to optimize, then the next data is ready to be processed to get the results of detection accuracy. If no error occurs, then the output data can be directly analyzed, but if an error occurs then it must be repeated again from the wrapper feature selection stage.

A. Metric

The metrics that will be the performance comparison points between the proposed method and the comparison methods are:

1. 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 usedTP + TN

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

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

2. Run Time

The run time between using wrapper method and without wrapper will be compared.

IV. Results And Discussion

A. Result

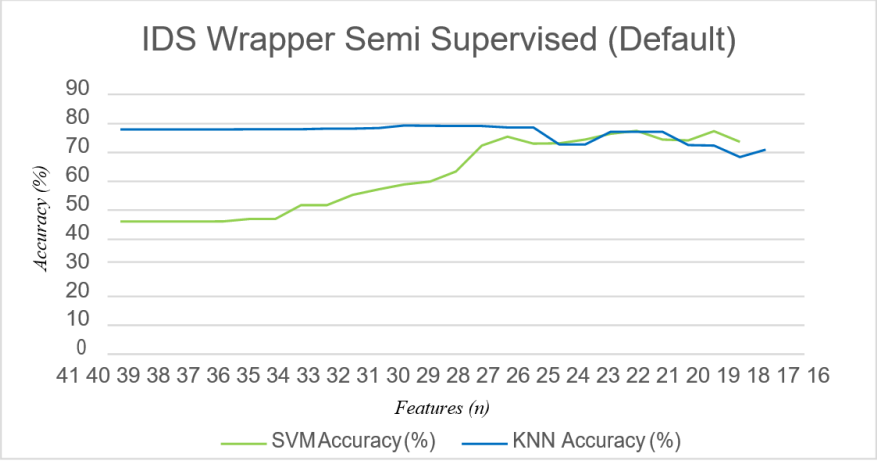

Fig 2. IDS *Wrapper Semi Supervised (Default)*

Based on the graph above, with the default parameter the accuracy value in SVM-RBF looks stagnant up to the number of feature 37th. Accuracy values are increasing as more features are selected up to the 25th feature count. After that there is fluctuation of accuracy value up to the number of feature 16th. While the KNN looks stagnant until the number of features to-31. After that, the accuracy value tends to increase slightly to the number of features to-27. And after that fluctuations occur accuracy value but tends to fall until the number of features to- 17, and increased again when the 16th feature.

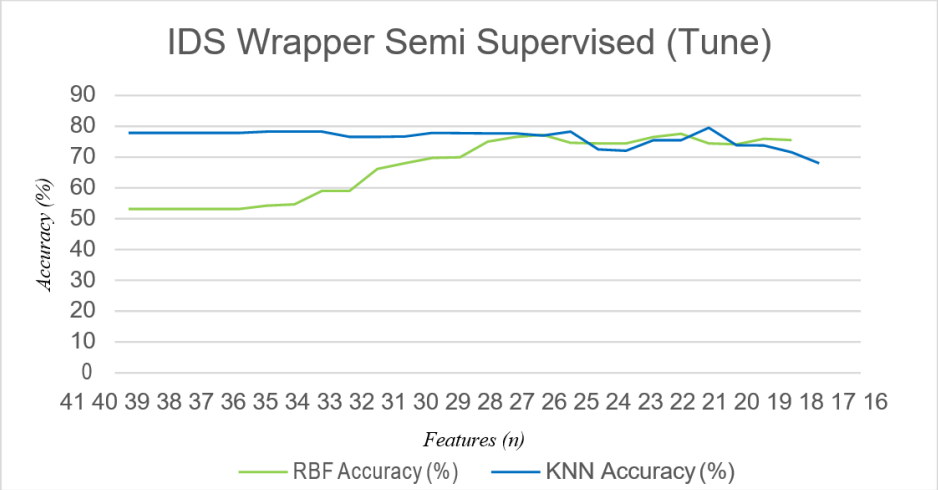

Fig 3. IDS *Wrapper Semi Supervised (Optimized)*

Based on the graph above, with the optimized parameters (tuning) the accuracy value in SVM-RBF looks stagnant up to the number of feature 37th. Accuracy values are increasing as more features are selected up to the 25th feature count. After that there is fluctuation of the value of accuracy until the number of features to 16 but tend to be adjacent. While the KNN tends to stagnate until the number of features to-34. After that, the accuracy value looks menunrun but tends to stagnate until the number of features to-25. Thereafter there was a decrease to the number of features to 23 and increased to the number of features to-20. Afterwards the accuracy value tends to decrease. To see the full test results can be seen on the attachment sheet.

B. Discussion

The first test is parameter testing on semi supervised IDS without using wrapper to get best parameter tuning on accuracy result. The result of semi supervised IDS testing with default setting is SVM-RBF (gamma = 20, n_neighbors = 7) i.e 46.027% accuracy with detection time 56.189s. Gamma is a parameter in SVM-RBF. The greater the value of gamma, the farther out of reach and the smaller the value of the variance, and vice versa. Based on the test results, SVM-RBF with gamma = 10 recorded the highest accuracy result 53.178% with time 60.120s. The smaller the gamma value the higher the accuracy. While in KNN, the parameters that influenced the expected accuracy result is n_neighbors. Based on the test results, KNN with n_neighbors = 7 recorded the highest yield of 77.966% with time 6.835s. From the research, it was concluded that with parameter gamma = 10 and n_neighbors= 7 got the highest accuracy result. This is attempted to apply to each classification by setting the same parameters, but there is no increase in accuracy and only little impact on the detection time.

The test results on the wrapper generally show that the larger the reduced feature the longer processing time required. The application of wrapper on semi supervised IDS also takes longer processing time when compared to without wrapper. In the result of accuracy detection by using the default parameter that is RBF with gamma = 20, wrapper record highest accuracy is 77.478% with time 101.375s on 20 features. If the parameter in the classification is changed to KNN, the highest accuracy is 79.323% with 31.55s time on 30 features. From these results, the classification setting affects the results of the wrapper accuracy. To improve the accuracy result, the

wrapper tuning the random_state parameter to get the highest result on SVM-RBF with 81.275% accuracy with 91,595s detection time. whereas in KNN classification no increase in accuracy.

In addition to the classification, the number of selection features also affect the results of the wrapper accuracy. From the test table on the attachment sheet, the feature on NSL-KDD has been ranked by rank. Can be analyzed that feature protocol_type to dst_host_srv_rerror_rate is a feature that is considered important by the wrapper. While is_host_login feature is a feature that is considered least important by the wrapper.

V. Conclusion

Based on the results of the experiments conducted on the research, the wrapper feature selection resulted in higher attack detection accuracy than the method without wrapper. By default parameter setting the highest accuracy result on wrapper is 77.478% with detection time 101.375s, while method without wrapper has accuracy 46.054% with detection time 57.253s. And after the optimization, the detection accuracy on the wrapper reached 81.275% with a detection time of 91.595s. From the results of the research, wrapper method provides a significant increase in accuracy although it requires a longer processing time. In addition, there are supporting factors to improve the accuracy of the number of features selection, tuning on the classification, and tuning on the wrapper. It can be concluded that the application of wrapper on semi supervised IDS can increase accuracy significantly.

Acknowledgment

The authors would to 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 to my colleague friend graham who help us while doing this research. Finally, with all of our respect we hope this research can give the real impact in 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.