A Combination of PSO-Based Feature Selection and Tree-Based Classifiers Ensemble for Intrusion Detection Systems

Bayu Adhi Tama and Kyung Hyune Rhee

Abstract Due to the numerous attacks over the Internet, several early detection systems have been developed to prevent the network from huge losses. Data mining, soft computing, and machine learning are employed to classify historical network traffic whether anomaly or normal. This paper presents the experimental result of network anomaly detection using particle swarm optimization (PSO) for attribute selection and the ensemble of tree-based classifiers (C4.5, Random Forest, and CART) for classification task. Proposed detection model shows the promising result with detection accuracy and lower positive rate compared to existing ensemble techniques.

Keywords Particle swarm optimization · Anomaly detection · Ensemble of tree-based classifiers

1 Introduction

With the rapid growth of computer networks, the number of users connected to the Internet has increased year by year. Severe disasters might be risen due to the excessive escalation of malicious intrusion or attack over the Internet. Therefore, the need for providing secure and safe security systems through the use of intrusion detection systems (IDS), encryption, or firewall is required. An IDS plays a vital role to analyze the network events occurring in a computer networks for indication of intrusion presence.

Intrusion aims at attempting to violate computer security policies such as confidentiality, integrity, and availability [1]. To date, significant research concern

B.A. Tama · K.H. Rhee(✉)
Lab. of Information Security and Internet Applications,
IT Convergence and Application Engineering,
Pukyong National University (PKNU), Busan, South Korea
e-mail: {bayuat,khrhee}@pknu.ac.kr

D.-S. Park et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*,
Lecture Notes in Electrical Engineering 373,
DOI: 10.1007/978-981-10-0281-6_71

in information security is intrusion detection and prevention. Intrusion detection can be considered as a classification analysis which is given to computer network traffic whether as normal or anomaly [2].

A first IDS was proposed by Denning [3], since then, numerous detection techniques including statistical methods, machine learning and data mining have been deployed in order to improve their performance. However, the deployment of efficient model to identify such malicious activity is still challenging task. An IDS must not have high computational burden and can perform intelligently so as to recognize previously unknown attacks. Specifically, an IDS must meet a low false positive rate and high detection rate [4].

In the recent work, a combination of PSO-based feature selection technique and multiple of tree-based classifiers system are proposed. PSO [5] is chosen to reduce computational cost since its capability to automatically search good features. We also consider the fusion of tree-based classifiers such as C4.5 [6], Random Forest [7], and CART [8] for classification analysis in order to increase detection accuracy. The performance result of proposed model is compared with the aforementioned base classifiers as single classifier and the state-of-the art ensemble techniques such as Bagging [9], Real Adaboost [10], MultiBoost [11], and Rotation Forest [12].

The objective of this paper are as follows: firstly, to select the most relevant features for intrusion detection systems using PSO and correlation-based feature selection (PSO-CFS); and secondly, to introduce the fusion of tree-based classifiers to maximizing the classification accuracy.

2 Related Work

Biology-inspired methods have tremendous impact to design computer security systems. They have developed novel and effective protection mechanism. Due to the increased deployment and widespread use of computer systems, traditional approaches often suffer from scalability problems to cope with [13]. Thus, it is important to consider biologically systems as sources of inspiration when designing new approaches [14].

PSO as one of many existing biology-inspired methods have been widely applied in IDS. It has been adopted for core functionality of IDS such as classification task or for secondary functions such as feature selection [4]. For instance, Zainal et al. [15] proposed the integration of rough set theory and particle swarm (Rough-DPSO) for feature selection process in IDS. From the experiment, proposed method offers better representation of data and they are robust. The most recent research regarding the use of PSO for feature selection in intrusion detection is a method called dynamic swarm based rough set (IDS-RS) [16]. IDS-RS is proposed to select the most relevant features that can represent the pattern of the network traffic.

2.1 PSO and Correlation-Based Feature Selection

PSO firstly proposed by Kennedy and Eberhart [17], is one of computation technique which is inspired by behavior of flying birds and their means of

information exchange to solve the problems. Each particle in the swarm represents possible solution. A number of particle is located in the hyperspace, which has random position φ_i and velocity ϑ_i. The basic update rule for the position and the speed is depicted in Eq. (1) and (2), respectively.

$$\varphi_i(t+1) = \varphi_i + \vartheta_i(t+1) \tag{1}$$

$$v_i(t+1) = \omega\vartheta_i(t) + c_1 r_1(p_i - x_i) + c_2 r_2(g - x_i) \tag{2}$$

Where ω denotes inertia weight constant, c_1 and c_2 denotes cognitive and social learning constant, respectively, r_1 and r_2 represent random number, p_i is personal best position of particle i, and finally, g is global best position among all particles in the swarm.

Correlation-based feature selection (CFS) is one of leading subset selection method in machine learning and pattern recognition [18]. CFS uses entropy and information gain theory to measure the uncertainty or unpredictability of a system. The lack of computation using information gain is symmetrical uncertainty and biased of feature with more values. Hence, CFS adopts a coefficient to compensate information gain's bias toward attribute with more values and to normalize its value to the range $[0,1]$.

In this paper, the integration of PSO and CFS is employed. An open source data mining tool, Weka, allows us to combine PSO and CFS as search and evaluation method, respectively. We consider to compare the number of selected features by varying the number of particle and its influence to the performance of classifier.

2.2 Fusion of Tree-Based Classifiers

Nowadays, the fusion of several base classifiers in parallel has been widely applied in many applications. Parallel approach organizes classifiers in parallel. All classifiers are applied for the same input in parallel, and then the result from each classifier is then combined to yield the final output. Moreover, in parallel approach, a combination rule is needed in order to incorporate the output of each classifiers. Once the base classifiers have been trained, a classifier fusion is formed by the rule of voting. In this current work, we consider to compare the accuracy of classifiers fusion using majority voting [19] and average of probabilities rule [20].

Our approach is based on the hypothesis that the use of classifiers fusion, an accurate detection can be obtained. Our base classifiers are C4.5, Random Forest, and CART. The selection of base classifiers, although we could choose other classifiers, is based on the fact that these classifiers are widely applied in many today's applications and show successful performance.

3 Experimental Setup

In this section, the scheme for intrusion detection using PSO-based feature selection and multiple classifier systems is presented. Firstly, we present a

framework of intrusion detection using multiple classifiers systems as depicted in Figure 1. Then, each part of the framework such as dataset description, parameter for feature selection process, and the strategy for combining multiple tree-based classifiers will be briefly discussed.

Our experiments were performed using NSL-KDD dataset [21]. It consists of selected records of older and well-known dataset, KDD Cup. NSL-KDD possesses 41 attributes plus one class label. It has 12973 records with 53.3% of normal class and anomaly class (represents 23 attacks) for the rest. All attributes were labeled from A to AO. Dataset is divided into 2 parts. One part which consists of two-third (66%) of dataset will be used for training, and the rest will be used for evaluation.

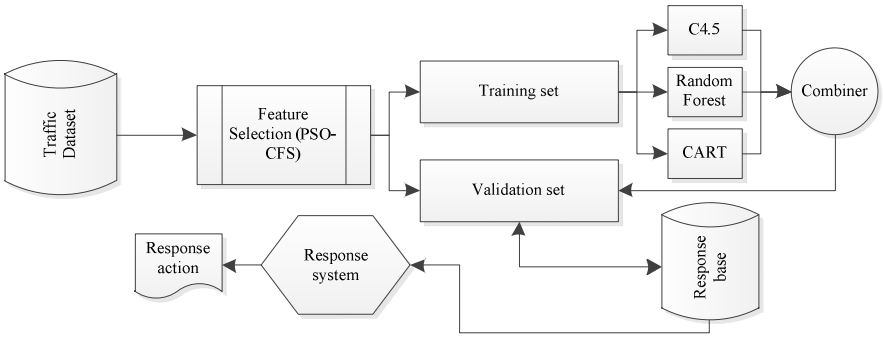

Fig. 1 Framework of intrusion detection systems

The parameter for the feature selection process using PSO-CFS the learning parameter of base classifiers is shown in Table 1. A different number of particle for feature selection are 50, 100, and 200 are denoted by PSO-50, PSO-100, and PSO-200, respectively. We considered the same parameters for base classifiers either as part of MCS or as an individual classifier. We ran the experiments using Weka, running on a system with an Intel Core i5 3.65GHz, 16GB RAM, and Windows 7 Professional.

The performance of classifiers are measured by accuracy and false positive rate (FPR). Accuracy represents the percentage of correctly classified of samples for different number of attributes, whilst FPR denotes the number of incorrectly classified of samples as belonging to positive class.

Table 1 Parameter for feature selection and learning process

PSO-CFS	C4.5	Random Forest	CART
Number of particles: 50, 100, and 200	Confidence factor, C = 0.25	Number of trees: 100	Heuristic process
ω = 0.33	Number of folds = 3	Number of slots = 1	Number of pruning: 5
c_1, c_2 = 0.34	Min. instance per leaf = 2		Pruning
	Pruning		

4 Result and Discussion

Table 2 summarizes our experimental results. Our proposed scheme with respect to classifiers ensemble using tree-based classifiers performed better than single classifier and existing ensemble classifiers. Based on the feature selection experiment by varying the number of particles, as the number of particle increases, the number of selected features continue to decreases significantly. Nevertheless, the fewer of selected features, the lower performance of classifier has.

Moreover, proposed feature selection scheme PSO-50 with average of probability voting ensemble scheme shows higher accuracy rate than other classifiers. It can be said that in the future, ensemble of tree-based classifiers might become a promising solution to detect anomaly in computer network. With reference to single classifier, RF always performs better than other classifiers with 99.78%, 99.67%, and 99.43% of predictive accuracy for PSO-50, PSO-100, and PSO-200, respectively. Surprisingly, among the ensemble technique, Rotation Forest with C4.5 as base

Table 2 Cross comparison results

Method	Selected Features	Accuracy (%)	FPR
PSO-50			
C4.5	D, E, F, L, Z, AC, AD, AG, AK, AL, AM	99.71	0.003
RF		99.78	0.002
CART		99.72	0.003
Bagging-C4.5		99.76	0.002
Real Adaboost-C4.5		99.77	0.002
Multiboost-C4.5		99.78	0.002
Rotation Forest-C4.5		98.98	0.011
Maj. Voting (C4.5+RF+CART)		99.76	0.002
Average of Prob. (C4.5+RF+CART)		**99.80**	**0.002**
PSO-100			
C4.5	D, E, F, L, Z, AC, AD, AK, AM	99.62	0.004
RF		99.67	0.004
CART		99.60	0.004
Bagging-C4.5		99.62	0.004
Real Adaboost-C4.5		99.64	0.004
Multiboost-C4.5		99.64	0.004
Rotation Forest-C4.5		98.39	0.018
Maj. Voting (C4.5+RF+CART)		99.64	0.004
Average of Prob. (C4.5+RF+CART)		**99.76**	**0.002**
PSO-200			
C4.5	D, E, F, L, Z, AD	99.39	0.007
RF		99.43	0.006
CART		99.35	0.007
Bagging-C4.5		99.38	0.007
Real Adaboost-C4.5		99.40	0.007
Multiboost-C4.5		99.38	0.007
Rotation Forest-C4.5		98.12	0.019
Maj. Voting (C4.5+RF+CART)		**99.44**	**0.006**
Average of Prob. (C4.5+RF+CART)		99.39	0.007

classifier continue to show unsatisfactory performance compared to other ensemble technique. Finally, a good performance could not be obtained when we employed a well-known decision tree method C4.5 as single classifier, yet C4.5 tends to show good performance when we incorporate it in ensemble technique.

5 Conclusion

In this paper, the performance of particle swarm optimization feature selection based and classifiers ensemble are thoroughly studied. The classifiers ensemble was built by tree-based classifiers such as C4.5, Random Forest, and CART, while feature selection was carried out by varying the number of particles. The proposed scheme showed good performance compared to other ensemble techniques. An incorporation of fifty particles of PSO and an average probability voting rule gave us promising performance with 99.8% of accuracy. As future work we will study the performance of classifiers fusion by combining other machine learning technique with different dataset.

Acknowledgement This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2014R1A2A1A11052981).

References

1. Liao, H., Lin, C., Lin, Y., Tung, K.: Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications **36**(1), 16–24 (2013)
2. Catania, C., Garino, C.: Automatic network intrusion detection: Current techniques and open issues. Computers & Electrical Engineering **38**, 1062–1072 (2012)
3. Denning, D.: An intrusion-detection model. IEEE Transactions on Software Engineering **2**, 222–232 (1987)
4. Kolias, C., Kambourakis, G., Maragoudakis, M.: Swarm intelligence in intrusion detection: A survey. Computers & Security **30**(8), 625–642 (2011)
5. Moraglio, A., Di Chio, C., Poli, R.: Geometric particle swarm optimisation. In: Genetic Programming. LNCS, vol. 4445. Springer (2007)
6. Quinlan, J.: C4. 5: programs for machine learning. Elsevier (1993)
7. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001)
8. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and regression trees. CRC Press (1984)
9. Breiman, L.: Bagging predictors. Machine Learning **24**(2), 123–140 (1996)
10. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Annals of Statistics **95**(2), 337–407 (2000)
11. Webb, G.: MultiBoosting: A Technique for Combining Boosting and Wagging. Machine Learning **40**(2), 159–196 (2000)
12. Rodriguez, J., Kuncheva, L., Alonso, C.: Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(10), 1619–1630 (2006)

13. Williamson, M: Biologically Inspired Approaches to Computer Security. Technical Report, HP Laboratories, Bristol (2002)
14. Twycross, J., Aickelin, U.: An immune-inspired approach to anomaly detection. In: Handbook of Research on Information Security and Assurance. IGI Global (2008)
15. Zainal, A., Maarof, M., Shamsuddin, S.: Feature selection using rough-DPSO in anomaly intrusion detection. In: Computational Science and Its Applications. LNCS, vol. 4705, pp. 512–524 (2007)
16. Chung, Y., Wahid, N.: A hybrid network intrusion detection system using simplified swarm optimization (SSO). Applied Soft Computing **12**(9), 3014–3022 (2012)
17. Kennedy, J., Eberhart, R.: A discrete binary version of the particle swarm algorithm. In: IEEE International Conference on Systems, Man, and Cybernetics, pp. 4104–4108 (1997)
18. Hall, M.: Correlation-based feature selection for machine learning. The University of Waikato, Hamilton (1999)
19. Kuncheva, L.: Combining pattern classifiers: methods and algorithms. John Wiley and Sons (2004)
20. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(3), 226–239 (1998)
21. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.: A detailed analysis of the KDD CUP 99 data set. In: Second IEEE Symposium on Computational Intelligence for Security and Defence Applications (2009)