Optimal feature selection for firewall log analysis using Machine learning and Hybrid Metaheuristic algorithms

1st Seungwoo Han
Kwangwoon University
Seoul, Korea
seungwoohan0108@gmail.com

2nd Gil Hong
Sahmyook University
Seoul, Korea
wideroad95@nate.com

3rd Jewan Kim
Sahmyook University
Seoul, Korea
wpdhks6@gmail.com

4th Jeuk Yu
Sahmyook University
Seoul, Korea
jipsina71@gmail.com

5th Sangjun Lee
Sahmyook University
Seoul, Korea
tjdtlf1234@naver.com

6th Byeongok Cho
Sahmyook University
Seoul, Korea
winjbo1021@gmail.com

7th Jusung Jeon
Sahmyook University
Seoul, Korea
gggg1371@naver.com

***Abstract*—Firewall log classification is important to monitor network traffic. Most firewall log classification via machine learning has shown good result by network-related features and classifiers. However, feature with many dimensions take a lot of time to do classification. In this paper, we applied a method of feature selection using optimized bee swarm optimization with reinforcement learning. We evaluated average performance by accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1 score in 5-stratified folds using a random forest, k-nearest neighbor, and naïve bayes classifier. As a results, it could be applied for an automatic firewall log analysis system.**

Keywords—bee swarm optimization, reinforcement learning

I. Introduction

Firewalls of computers are used to ensure that the network is functioning properly and safely. In particular, as the security of personal information [1] becomes more important and robust, it needs to be conscious to protect networks. Firewall logs are key evidence to identify intruder attacks, including insider and outsider threats [2]. In addition to the existing traditional classification methods, with the development of machine learning and deep learning, a study on log classification and intrusion prevention using it has been conducted [3–8]. Log analysis and intrusion detection defense should attempt to classify efficiently with fewer parameters for quick response. At this time, the number of features used for classification or regression is based on the researcher's experience. For this reason, research is also being conducted to intensively select the optimal feature and parameter via information gain, and genetic algorithms, and reinforcement learning [9–12]. In this paper, we performed optimal feature search using the optimized bee swarm optimization algorithm along with reinforcement learning [12].

II. Experiment

A. Data acquisition

We used Internet Firewall Data Data Set [13] in UCI Machine Learning Repository. The data have 11 features and 4 labels. Total data points are 65532. Data profile and feature information are shown in Table 1 and Table 2. Four label is allow, deny, drop, reset-both, respectively.

Table 1: Profile of Dataset

Name	Data points
Allow	37640
Deny	14987
Drop	12851
Reset-both	54
Total	66532

Table 2: Feature information

Name			
bytes	bytes received	bytes sent	destination port
elapsed time	NAT destination port	packets	packets received
NAT source port	packets sent	source port	

B. Methods

1. Bee swarm optimization algorithm

bee swarm optimization algorithm (BSO) is an algorithm that is inspired by the social behavior of bees. Each bee is an object working together to solve the optimization problem, and they search for the fitness function using a feature combination in iterations. The fitness function of this research is set to average accuracy.

2. Reinforcement learning

Reinforcement learning (RL) refers to an algorithm in which an agent defined in the environment recognizes the current state and finds an action that maximizes the reward among actions. The RL algorithm applied in this paper is Q-learning [14]. Local search and experience of bee replace Q-learning. In this process, the reward is given differently depending on the accuracy of the current and next states. The reward-setting condition is shown Table 3.

Table 3: Reward-setting condition

Condition	Reward
if the next state accuracy is higher than the current state	next state accuracy
if the current state accuracy is high	next state accuracy - current state accuracy
if the number of features in the current state is greater than the number of features in the next state	1/4 * next state accuracy
if the number of features in the current state is less than the number of features in the next state	-1/4 *next state accuracy

As a result, the agent tries to get the best accuracy while getting fewer features. Moreover, to reduce the space in the search space, we applied the XOR operation on the best solution and the current state solution.

3. Classifier

In this paper, We compared performance using several classifier; random forest (RF), k-nearest neighbor (KNN), and naïve baye (Bernoulli NB).

III. HYPER-PARAMETER SETTING

Table 4 shows the hyper-parameters applied to this experiment.

Table 4: Hyper-parameter lists

Name	Parameter	Description
Flip	5	To calculate Search Region space
Max Chance	3	Number of chances to escape local minima
Number of Bee	3	Worker to solve
Max iteration	3	Number of total iteration
Local iteration	2	Number of iteration in local search
alpha	0.8	Learning rate of reinforcement learning
gamma	0.3	Discount factor of reinforcement learning
epsilon	0.05	Probability of doing random actions

The machine learning classifier was used python library called scikit-learn (version 0.24.2). hyper-parameter of classifier was used default parameter.

IV. RESULT

The optimal combination list of features in training set selected by the feature selection method is shown Table 5.

Table 5: Optimal feature lists

Classifiers	Optimal feature combination lists
RF	Destination port NAT Source Port NAT Destination Port Bytes Bytes Received Packets Elapsed Time Packets sent Packets received
KNN	Destination Port NAT Destination Port Bytes Bytes Sent Packets Elapsed Time
Bernoulli NB	Source Port Bytes Bytes Sent Elapsed Time Packets received

And we evaluate accuracy performance each classifier. Table 6 show the best accuracy among combination lists each classifiers.

Table 6: Best accuracy (%)

Classifiers	Best accuracy among combination lists
RF	**99.87**
KNN	99.78
Bernoulli NB	80.22

And we measure random forest performance by average accuracy and average macro-averaged precision (Macro-precision), average macro-averaged recall (Macro-recall), average macro-averaged F1 (Macro-F1) score. We compared the optimal selected feature results with the case of applying all features using random forest. The results are shown in Table 7.

Table 7: Comparing all and optimal features using RF (%)

Feature selected	Average Accuracy	Macro-precision	Macro-recall	Macro-F1
All	99.80	93.06	81.71	84.22
Optimal	**99.87**	**96.62**	**88.73**	**91.27**

V. CONCLUSION

In this paper, we classified firewall logs using optimal feature via BSO with reinforcement learning feature selection method. The results of using optimal features outperformed using all features and it could be applied to a firewall log analysis.

REFERENCES

[1] J. A. Castañeda and F. J. Montoro, "The effect of Internet general privacy concern on customer behavior", InElectronic Commerce Research, vol. 7, no. 2, pp. 117–141, 2007.

[2] N. K. Singh, D. S. Tomar, and B. N. Roy "An approach to understand the end user behavior through log analysis",InInternational Journal of Computer Applications, vol. 5, no. 11, pp. 27–34, 2010.

[3] S. Allagi and R. Rachh, "Analysis of Network log data using Machine Learning", In2019 IEEE 5th InternationalConference for Convergence in Technology (I2CT), pp. 1–3, 2019.

[4] F. Ertam and M. Kaya, "Classification of firewall log files with multiclass support vector machine", In2018 6thInternational Symposium on Digital Forensic and Security (ISDFS), pp. 1–4, 2018.

[5] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", In IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.

[6] D. C. Le, N. Zincir-Heywood, and M. I. Heywood, "Analyzing Data Granularity Levels for Insider Threat Detection Using Machine Learning", InIEEE Transactions on Network and Service Management, vol. 17, no. 1, pp. 30–44,Mar. 2020.

[7] A. Javaid, Q. Niyaz, W. Sun, and M. Alam "A deep learning approach for network intrusion detection system", InProceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies(formerly BIONETICS), pp. 21–26, 2016.

[8] D. Sharma, V. Wason, and P. Johri, "Optimized classification of firewall log data using heterogeneous ensemble techniques," in 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 368–372. 2021.

[9] S. Lei, "A Feature Selection Method Based on Information Gain and Genetic Algorithm", In2012 International Conference on Computer Science and Electronics Engineering, pp. 355–358, 2012.

[10] J. Janisch, T. Pevný, and V. Lisý, "Classification with Costly Features Using Deep Reinforcement Learning", InProceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3959–3966, 2019.

[11] W. Fan, K. Liu, H. Liu, P. Wang, Y. Ge, and Y. Fu, "AutoFS: Automated feature selection via diversity-aware interactive reinforcement learning," in 2020 IEEE International Conference on Data Mining (ICDM), pp. 1008–1013. 2020.

[12] S. Sadeg, L. Hamdad, A. R. Remache, M. N. Karech, K. Benatchba, and Z. Habbas, "QBSO-FS: A ReinforcementLearning Based Bee Swarm Optimization Metaheuristic for Feature Selection", InInternational Work-Conferenceon Artificial Neural Networks, vol. 33, pp. 785-796, 2019.

[13] A. Frank, and A. Asuncion, "UCI Machine Learning Repository http://archive.ics.uci.edu/ml", Irvine, CA:University of California, School of Information and Computer Science. 2010.

[14] C. J. Watkins and P. Dayan, "Q-learning",Machine learning, vol. 8, no. 3-4, pp. 279–292, 1992.