Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM

ISSN: 2788-7421

Rao-SVM Machine Learning Algorithm for Intrusion Detection System

Shamis N. Abd[1], Mohammad Alsajri[2*], Hind Raad Ibraheem[3]

[1]Department of Computer Science, Al-Salam Uiversity College

[2]Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Malaysia

[3]Department of Computer Science, Al-Salam Uiversity College

*Corresponding Author: Mohammad Alsajri

DOI: https://doi.org/10.52866/ijcsm.2019.01.01.004
Received December 2019; Accepted December 2019; Available online January 2020

ABSTRACT: Most of the intrusion detection systems are developed based on optimization algorithms as a result of the increase in audit data features; optimization algorithms are also considered for IDS due to the decline in the performance of the human-based methods in terms of their training time and classification accuracy. This article presents the development of an improved intrusion detection method for binary classification. In the proposed IDS, Rao Optimization Algorithm, Support Vector Machine (SVM), Extreme Learning Machine (ELM), and Logistic Regression (LR) (feature selection and weighting) were combined with NTLBO algorithm with supervised ML techniques (for feature subset selection (FSS). Being that feature subset selection is considered a multi-objective optimization problem, this study proposed the Rao-SVM as an FSS mechanism; its algorithm-specific and parameterless concept was also explored. The prominent intrusion machine-learning dataset, UNSW-NB15, was used for the experiments and the results showed that Rao-SVM reached 92.5% accuracy on the UNSW-NB15 dataset.

Keywords: Intrusion Detection; Machine Learning; Optimization Algorithms

1. INTRODUCTION

Serious confidentiality, privacy and security issues are being associated with the use of the internet in recent times due to the processes involved in data transformation and transmission across the internet. This has necessitated much effort towards improving the privacy and security of computer systems; however, these issues are yet to be properly addressed as there is currently no completely secure system in the world. Furthermore, there are numerous types of network attacks [1] which evolves when new attack signatures are added to the signature database. The emergence of new attack signatures has driven the urge to develop novel systems for the detection of such attacks as they emerge. Intrusion detection system is one the tools used to detect these new attacks as they can monitor and detect a range of network systems, information systems, and cloud computing system. The work of the IDS is to monitor a system and detect the presence of attacks that are aimed at attacking the availability, integrity, and confidentiality of a system. This paper review the existing work, methods and techniques in IDS, section II give an overview about IDS, then followed by brief description about the main types of IDS and the techniques used in detection explained in II.A and II.B, section III state the existing challenges exist

*Corresponding author: pcc17021@stdmail.ump.edu.my in modern IDS, in section IV the most used ML algorithms in the IDS are reviewed in details the main weaknesses and strengths of each algorithm is given in section IV.F. Section V explain two types of optimization algorithms parameters containing and parameters less algorithms. Finally, conclusion on what have been done given in section VI.

2. RELATED WORKS

Intrusion detection systems are deployed on network systems for the monitoring of such systems for various sources of intrusion. The existing IDSs fall into two categories - host-based and network-based IDS [2] [3]. For the network-based IDS (NIDSs), they can identify network intrusion through the analysis of specific network patterns, but for the host-based IDS (HIDS), their work is mainly to detects intruders in individual hosts. NIDSs are deployed to scan a packet sniffer or network switch output; a sniffer is a program that reads the raw packets of a local network segment. A peculiar feature of NIDSs is that they can detect and identify attacks that may be missed by the HIDSs because the HIDSs are not designed to see packet headers; hence, they cannot detect some types of network attacks. NIDSs, for instance, can detect only numerous IP-based DoS attacks because they can see the packet headers as they travel through the monitored network. On the other hand, HIDSs require different OS to operate properly unlike NIDSs which do not require any OS of the host as a source of attack identification. IODSs can also be categorized into misuse detection systems and anomaly detection systems [3]. Hybrid IDSs are developed as a combination of both HIDS and NIDS [4]. The detection mechanism used in IDS are three main types which is: statistical method, Machine Learning, and Data-Mining methods, Figure 1 summarize the IDS.

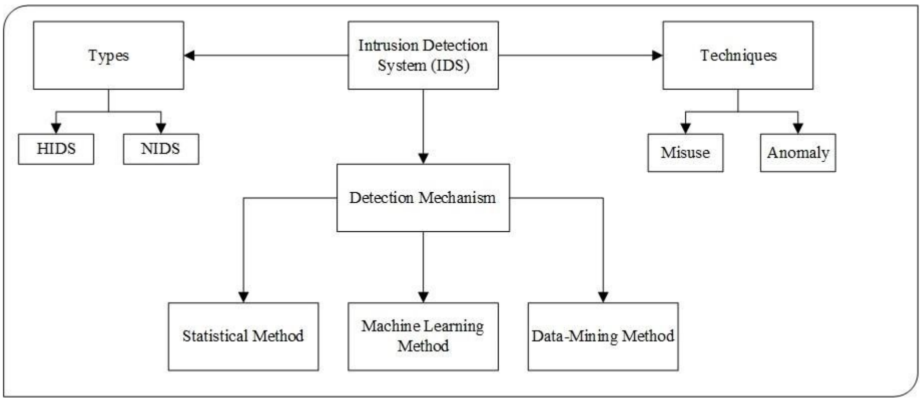

FIGURE 1. **IDS overview**

2.1 MISUSE DETECTION

Detection of abuse detects intrusions through looking for known patterns of attacks. This strategy is employed by current commercial NIDSs. A downside to preventing misuse is that it cannot spot unknown attacks. Specific methods, such as expert networks, signature analysis, state-transition analysis , and data mining, have been used to identify violations. To describe intrusions, the expert system uses a set of rules [2]. Audit events are converted into facts in the expert framework which bear their semantic significance. Then, using certain rules and evidence, an inference engine will draw conclusions. State transition analysis aims at the identification of attacks based on a set of goals and transitions using state transition.digrams An intrusion system detects events that trigger an intrusion state. Signature analysis is used to describe a new attack based on signatures that are already contained in audit trail [2], where network patterns that matches the signatures in the database are considered an intrusion. Numerous works are available on the use of data mining-based method for ID. Data mining is a technique for the extraction of important and previously unnoticed patterns from large datasets; these patterns may be defined as decision chains, rules, neural networks, or instance-based instances. The common DM algorithms for misuse detection include Mining Audit Data for Automatic Models for Intrusion Detection (MADAM ID) [4], Intrusion Detection Utilizing Data Mining Techniques (IDDM), and Audit Data Review and Processing (ADAM) [2]; these models are based on the Association rules algorithm [5]. The IDS performance is enhanced with the neural network algorithm [6]

2.2 ANOMALY DETECTION

Because detection of misuse cannot identify unknown threats, detection of abnormalities is used to counter this shortcoming. Different approaches to anomaly detection Clustering, Classification, etc. were proposed and implemented [7]. Supervised detection of anomalies makes use of attack-free training data for the creation of the normal traffic patterns; any deviation from the established normal traffic pattern is detected as an intrusion. ADAM [8] creates the normal traffic pattern from attack-free training dataset and describes the profile as a set of association rules. It performs real-time detection of suspicious connections based on the profile. Other supervised methods, such as genetic algorithms [9], fuzzy data mining, SVM and neural networks [10, 11] are also used to detect anomaly. Supervised anomaly detection often requires the use of specialized structures and mathematical techniques [2]. The user profile is created using statistical methods based on several instances of normal behavior. New behaviors are then compared against the normal profiles, and deviations are detected as an intrusion. For the expert systems, they describe the normal user behavior using a set of rules; these rules are applied to detect intrusions.

3. THE PROPOSED METHOD

This study proposed the execution of the RAO algorithm at the FSS phase; the RAO was initialized by an initial population that was generated randomly; the population is made up of the Teacher and a set of Students which are considered a set of potential solutions. The features of the RAO were represented by incorporating the crossover and mutation operators of GA; the enables the representation of the RAO features as chromosomes. The crossover operator is used to update the chromosome. Each solution in each generation is considered an individual or a chromosome as represented in Figure 1. Selected features of a chromosome are marked 1 while non-selected ones are marked 0. The detail of the new method is shown in Algorithm 1 while the flow of the method is shown in Figure 4.

Algorithm 1: presents the details of the RAO-SVM algorithm.

i *Step 1 initialize the population randomly with each population having different set of features*

ii *Step 2 Based on the accuracy of the classification for each set of features, specify best and worst set (population)*

iii *Step 3 modify solutions based on the best and worst solutions and random interactions based on New_set= random_set crossover with (best_set crossover with worst_set)*

iv *Step 4 if the new set of features better than the old best set (in term of accuracy of classification) then keep the new set else keep the old set*

v *Step 5. Is the termination criteria satisfied or not, if yes report the best set of features, else go to step 3*

4. EXPERIMENTAL SETUP

The experimental scenario, problem instances, and the outcome of the experiments are all presented in this section. In this study, the experiments were performed intrusion dataset called UNSW-NB15 which were reduced, because of the focus on binary classification to accommodate only two classes (normal and intrusion). To ensure a better validation, K-fold validation was used, where the value of K is set to 10. [17]

5. DATASET

A research group created the UNSW-NB15 dataset at the Australian Centre for Cyber Security (ACCS) for testing the performance of new IDSs [10]. The dataset comprised of 100 GB of raw data captures with the IXIA Perfect Storm tool and a tcpdump tool; the raw data represents the simulated network traffic of both contemporary attack and modern normal behaviors. The raw data were captured over two simulation periods for a period of 16 h & 15 h. The dataset has a total size of 2.5 M records. A total number of 49 features were created using the Argus, Bro-IDS tools, and other 12 algorithms. The 49 features are categorized into 5 categories (flow features, content features, basic features, time features, & additional generated features) while 2 features served as a label (attack_cat that indicates the attack category & the normal state, as well as the label which takes the value 1 for attack and 0 for normal). The dataset has 9 attack categories which are Fuzzers, DoS, Exploits, Analysis, Backdoor, Shellcode, Worms, Generic, and Reconnaissance [16].

6. RESULTS

The tables below present the accuracy results for both datasets. The accuracy result of the UNSW-NB15 dataset is presented in Table 1.

Table 1. **Accuracy result of UNSW-NB15 dataset.**

Classifier	**Rao**	
	No. of features	Accuracy
LR	17	0.921
	18	0.923
SVM	16	0.922
	19	0.925
ELM	19	0.92
	20	0.9215

From Table 4, both RAO-SVM and RAO-SVM offered the same execution time for each ML technique. For each ML, the number of features, accuracy, and execution time were calculated. The numbers in red suggest the best results for both RAO-SVM and RAO-SVM. RAO-SVM consistently presented better accuracies as compared to RAO-SVM using the three ML techniques. It also presented better time accuracy using LR and SVM ML techniques. However, RAO-SVM provided a better execution time with ELM as compared to RAO-SVM.

7. CONCLUSIONS

This paper proposes a new (RAO-SVM) for feature subset selection problems in intrusion detection. The performance of the new algorithm was demonstrated to be superior to many other algorithms in FSS problems on two large intrusion datasets. The proposed RAO-SVM consistently presented better accuracy in the execution time. On the statistical tests (confusion matrix) applied to the RAO-SVM detection rate and error rate extracted from the confusion matrix, RAO-SVM showed a higher detection rate for the UNSW-NB15 dataset. It showed a low error rate for the two datasets. As a recommendation, the proposed RAO-SVM should be applied to multi-class classification problems, and more ML techniques could be used for evaluating its performance.

REFERENCES

[1] I. Aljarah and S. Ludwig, “Mapreduce intrusion detection system based on a particle swarm optimization clustering algorithm,” *In Proceedings of IEEE Congress on Evolutionary Computation Conference, Cancun*, pp. 955–962, 2013.

[2] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, “Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model,” *Journal of Computational Science*, vol. 25, pp. 152–160, 2018.

[3] B. Altay, T. Dokeroglu, and A. Cosar, “Context-sensitive and keyword density-based supervised machine learning techniques for malicious webpage detection,” *Soft Computing*, vol. 23, no. 4, pp. 4177–4191, 2019.

[4] M. Alsajri, M. A. Ismail, and S. Abdul-Baqi, “A review on the recent application of Jaya optimization algorithm,” *in 2018 1st Annual International Conference on Information and Sciences (AiCIS), Fallujah*, pp. 129–132, 2018.

[5] S. M. H. Bamakan, H. Wang, T. Yingjie, and Y. Shi, “An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization,” *Neurocomputing*, vol. 199, pp. 90–102, 2016.

[6] J. Cai, J. Luo, S. Wang, and Y. S, “Feature selection in machine learning: A new perspective,” *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[7] A. Chaudhary, V. Tiwari, and A. Kumar, “A novel intrusion detection system for ad hoc flooding attack using fuzzy logic in mobile ad hoc networks,” *International Conference on Recent Advances and Innovations in Engineering, 2014, Jaipur*, pp. 1–4, 2014.

[8] M. Črepinšek, S.-H. Liu, and L. Mernik, “A note on teaching–learning-based optimization algorithm,” *Information Sciences*, vol. 212, pp. 79–93, 2012.

[9] M. Dash and H. Liu, “Feature selection for classification,” *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[10] S. P. Das, N. S. Achary, and S. Padhy, “Novel hybrid SVM-TLBO forecasting model incorporating dimensionality reduction techniques,” *Applied Intelligence*, vol. 45, no. 4, pp. 1148–1165, 2016.

[11] S. Das and P. S, “A novel hybrid model using teaching-learning-based optimization and a support vector machine for commodity futures index forecasting,” *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 1, pp. 97–111, 2018.

[12] E. D. la Hoz, E. D. L. Hoz, A. Ortiz, J. Ortega, and B. Prieto, “PCA filtering and probabilistic SOM for network intrusion detection,” *Neurocomputing*, vol. 164, pp. 71–81, 2015.

[13] D. Ding, Q. Han, Y. Xiang, X. Ge, and X. Zhang, “A survey on security control and attack detection for industrial cyber-physical systems,” *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[14] T. Dokeroglu, “Hybrid teaching–learning-based optimization algorithms for the Quadratic Assignment Problem,” *Computers & Industrial Engineering*, vol. 85, pp. 86–101, 2015.

[15] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive learning algorithms and representations for text categorization,” *Proceedings of the seventh international conference on Information and knowledge management, Bethesda*, pp. 148–155, 1998.

[16] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, “A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems,” *Expert Systems with Applications*, vol. 42, no. 5, pp. 2670–2679, 2015.

[17] C. Guo, Y. Ping, N. Liu, and S.-S. Luo, “A two-level hybrid approach for intrusion detection,” *Neurocomputing*, vol. 214, pp. 391–400, 2016.

[18] H. E. Kiziloz, A. Deniz, T. Dokeroglu, and A. Cosar, “Novel multiobjective TLBO algorithms for the feature subset selection problem,” *Neurocomputing*, vol. 306, pp. 94–107, 2018.

[19] M. K. Khaleel, M. A. Ismail, U. Yunan, and S. Kasim, “Review on Intrusion Detection System Based on the Goal of the Detection System,” *International Journal of Integrated Engineering: Special Iss*, vol. 10, no. 6, pp. 197–202, 2018.

[20] W.-C. Lin, S.-W. Ke, and C.-F. Tsai, “CANN: An intrusion detection system based on combining cluster centers and nearest neighbors,” *Knowledge-based systems*, vol. 78, pp. 13–21, 2015.

[21] Y. Li, J.-L. Wang, Z.-H. Tian, T.-B. Lu, and C. Young, “Building lightweight intrusion detection system using wrapper-based feature selection mechanisms,” *Computers & Security*, vol. 28, no. 6, pp. 466–475, 2009.

[22] P. Louvieris, N. Clewley, and X. Liu, “Effects-based feature identification for network intrusion detection,” *Neurocomputing*, vol. 121, pp. 265–273, 2013.

[23] S. Mahdavifar and A. A. Ghorbani, “Application of deep learning to cybersecurity: A survey,” *Neurocomputing*, vol. 347, pp. 149–176, 2019.

[24] M. M. Hasan, R. Ahmed, M. Tapus, N. Shanan, and M, “A Focal load balancer based algorithm for task assignment in cloud environment,” *in 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Iasi*, pp. 1–4, 2018.