DOI:10.24996/ijs.2018.59.1B.23

ISSN: 0067-2904

Improving Detection Rate of the Network Intrusion Detection System Based on Wrapper Feature Selection Approach

Rana F. Najeeb*, Ban N. Dhannoon
Department of Computer Science, College of Science, AL-Nahrain University, Baghdad, Iraq.

Abstract

Regarding the security of computer systems, the intrusion detection systems (IDSs) are essential components for the detection of attacks at the early stage. They monitor and analyze network traffics, looking for abnormal behaviors or attack signatures to detect intrusions in real time. A major drawback of the IDS is their inability to provide adequate sensitivity and accuracy, coupled with their failure in processing enormous data. The issue of classification time is greatly reduced with the IDS through feature selection. In this paper, a new feature selection algorithm based on Firefly Algorithm (FA) is proposed. In addition, the naïve bayesian classifier is used to discriminate attack behaviour from normal behaviour in the network traffic. The FA selects the discriminating features from NSL-KDD dataset. The performance of the IDS in the detection of attacks was enhanced by the proposed model and compare with other models.

Keywords: Feature Selection, Firefly Algorithm, Intrusion Detection System, Naive Bayesian Classifier

تحسين معدل الكشف لنظام كشف التسلل للشبكة بالاعتماد على نهج الالتفاف لاختيار الميزة

رنا فارس نجيب*، بان نديم ذنون
قسم علوم الحاسبات،كلية العلوم،جامعة النهرين، بغداد،العراق

الخلاصة

وفيما يتعلق بأمن النظم الحاسوبية، فإن أنظمة كشف التسلل (IDSs) هي اجزاء أساسية للكشف عن الهجمات في المراحل المبكرة .وهم يرصدون ويحللون حركة الشبكة، ويبحثون عن سلوكيات غير طبيعية أو توقيعات هجومية للكشف عن عمليات الاقتحام في الوقت الفعلي .ومن العوائق الرئيسية لل(IDS) عدم قدرته على توفير حساسية ودقة كافية، إلى جانب فشلها في معالجة البيانات الهائلة .وتقلص إلى حد كبير مسألة وقت التصنيف في (IDS) من خلال اختيار الميزة. في هذا البحث يتم اقتراح خوارزمية اختيار ميزة جديدة اعتمادا على خوارزمية اليراع (FA). وبالاضافة الى ذلك، يتم استخدام (NBC) لتميز سلوك الهجوم من السلوك الطبيعي في حركة مرور الشبكة .(FA) تختار الميزات المتميزة من مجموعة بيانات NSL-KDD. تم تعزيز أداء (IDS) في الكشف عن الهجمات من قبل النموذج المقترح ومقارنتها مع نماذج أخرى لتطوير الأداء.

1. Introduction

In the world of data innovation today, computer security is a critical area. The preserving of total security for computer systems is a difficult task owing to the complex and diverse nature of the

*Email: stcs-rfn16@sc.nahrainuniv.edu.iq

computer substructures. The daily eruption of novel automated intrusion tools has given rise to the increase in the number of computer attacks. These attacks could be from an interior or lawful user from outside users .Intrusion detection methods (IDM) are systems for the identification and handling of malicious computer and network resource usage[1,2]. The IDM was developed to ensure the security of computer systems by discovering and informing un-authorized and abnormal situations as well as network security violation. There are two categories of intrusion detection techniques which are anomaly detection technique and misuse detection technique.

Misuse detection technique identifies intrusion by the matching of the attack features through the attacking feature library. Its speed of intrusion detection is high with a low chance of false alarms; though it does not identify non-designated attacks in the feature library, and cannot detect several new attacks.

In anomaly detection techniques, the usual features of user's behaviors are stored in the database, and the behavior of the current user is compared to those stored in the database. In the presence of a high rate of divergence, it can be said that an abnormal situation has occurred. It has the advantage of being comparatively irrelevant to the system and its strong versatility possibility of detecting novel attacks. However, because a complete description of the behavior of all the users in the system cannot be provided by conducting a normal contour, the user behaviors often change, and there is a high chance of false alarms [3, 4, 5].

During the development of an IDS that uses machine learning technique, one of the major factors to be considered is the design of appropriate features that represent activities and differentiate normal network usage from attacks [6, 7, 8]. Even though there are many features that have been proposed, the lack of publicly available datasets makes the objective evaluation and fair comparison of the proposed features difficult and similarly delays the systematic investigations into the effect of features on IDSs. To solve this problem, MIT Lincoln laboratory [9] formulated KDDCUP 1999 dataset, while Tavallaee et al. [10, 11] modified it to develop the NSL_KDD dataset. After these, the performance of IDS proposed by many researchers has been subsequently evaluated objectively using the KDD'99 and NSL_KDD datasets.

However, there are 41 features in the connection vectors processed from raw tcpdump data in the KDD'99 and NSL_KDD datasets. These datasets have therefore been considered as having too many features for real-time deployment in IDS. Researches on IDSs have recently overcome this problem by using only the feature parts that have attracted several researchers' attention. This is referred to as feature selection problems and has elicited the proposing of many feature selection methods. With feature selection, the computation time can be reduced, prediction performance can be improved, and the machine learning data or pattern recognition applications can be understood.

However, in many proposed feature selection methods, the central focus has been on the analysis of the individual feature relevance to the dataset using analysis measures such as dependency ratio, information gain or correlation coefficient [12, 13, 14]. In these methods, features are usually ranked in a suggested metrics order, and then removed based on the ranking results [15]. These approaches do not explicitly consider the combinatorial properties of features, despite the capability of the combinatorial properties of features to lead to emergent effects on the performance of IDSs; and this is a major drawback of these methods. In other words, important features with less individual information but highly informative when in combination with other features could be eliminated [16]. The reason for adopting the relevance of individual features to the data as a feature selection criteria in the proposed methods despite knowing the issues with these approaches is due to a large number of possible feature subsets. As the number of features in the KDDCUP data set is 41, the total number of feature subsets is up to $41-1$. Therefore, it is difficult to get the optimal feature subsets for IDSs based on the evaluation of the individual performance of the features rather than a collectively.

This study proposed a new feature selection method based on the binary Firefly Algorithm (FA) and Naïve Bayesian (NB) classifier. This paper is organized as follows: Section 2 presents the description and properties of the NSL_KDD dataset, while section 3 presents the details of the proposed feature selection algorithm. Section 4 comparers the performance of the selected feature subsets and 41 features. Finally, section 5 presents the major conclusions and recommendations for future research directions.

2. NSL_KDD Dataset

In this paper, the NSL_KDD dataset was used to assess the performance of the subsets selected by running the proposed algorithm. The original KDD'99 data set widely used for the evaluation of the performance of IDSs is made up of the test and train datasets, each with nearly 300 thousand and 5 million instances, respectively. Table-1 showed the instances of attack from 41 features that belong to one of the four forms of attack (*Denial of Service*, *User to Root*, *Remote to Local*, and *Probing Attacks).*

Table 1-Categories of attack reserved in KDD-Cup '99

Attack	Description
Denial of Service (Dos)	These attacks exhaust the network traffic or computing resources, denying the legitimate users of the services provided.
User to Root (U2r)	These attacks try to bypass the network after sniffing the ordinary users.
Remote to Local (R2l)	These attacks try exploiting the target server vulnerability to access the ordinary users.
Probing	These attacks collect network activity information in an attempt to avoid security management.

These 41 features are additionally grouped into 3 groups which are the basic, contents, and traffic features as shown in Table-2.

Table 2-Basic Three Groups of Features

Basic Features	Contents Features	Traffic Features
duration	Hot	count
protocol_type	num_failed_logins	serror_rate
service	logged_in	rerror_rate
src_bytes	num_compromised	same_srv_rate
dst_bytes	root_shell	diff_srv_rate
flag	su_attempted	srv_count
land	num_root	serror_rate
wrong fragment	num_file_creations	rerror_rate
urgent	num_shells	same_srv_rate
duration	num_access_files	srv_serror_rate
	num_outbound_cmds	srv_rerror_rate
	is_hot_login	
	is_guest_login	

There are some advantages of the NSL_KDD data set over the KDD'99 dataset even though it is a subset of KDD'99 dataset. At first, no redundant records exist in the NSL_KDD dataset as in KDD'99 train and test data set; so, there will be no bias in the learning algorithm based-IDS based towards more frequent records. Records from each attack category are adjusted based on its level of difficulty with respect to attack detection; making it easier to evaluate different detection methods with improved accuracy. Secondly, there is a reasonable amount of records in the NSL_KDD data set that made it possible to objectively compare different detection methods while avoiding the arbitrariness that occurs when using randomly selected data parts.

There are similar categories (4) of attacks in the NSL_KDD dataset as the KDD'99 dataset, with each data instance having 41 features. While in this paper is used the binary classes which include (Normal and Attack) only. This paper is, therefore, presenting a feature selection method for binary classification problem of IDS, employing a dataset of 10,000 records.

3. Firefly Algorithm (FA)

The FA is a nature-inspired biological global stochastic approach for optimization developed by Yang [17]. It is a meta-heuristic approach based on Firefly population, with each Firefly representing a potential search space solution. The FA copies the mating and light flash-based information exchange

mechanisms of Fireflies. In this section, the major attributes of Fireflies, the artificial FA, as well as the variants introduced to the basic algorithm already proposed were presented. Three idealized rules which describe the behavior of the artificial Fireflies were proposed by Yang [17] as:

1. Fireflies are unisex and can be attracted to each other irrespective of the sex.
2. The degree of attractiveness is related to the intensity of the emitted light; therefore, Fireflies with lights of lesser intensity will be made to move towards lights with higher intensities. Attractiveness decreases with increase in the distance between fireflies. They will randomly move when there is no brighter Firefly within the surrounding.
3. The brightness of the light from the Fireflies is a function of the landscape of the fitness function. The brightness can be proportional to the fitness function value of the maximization problem.

From these criteria, a summary of the basic steps of the FA can be presented as the pseudo-code illustrated in Figure-1.

Firefly Algorithm

Objective function $f(\mathbf{x})$, $\mathbf{x} = (x_1, ..., x_d)^T$
Generate initial population of fireflies $\mathbf{x}_i$ $(i = 1, 2, ..., n)$
Light intensity I_i *at* $\mathbf{x}_i$ *is determined by* $f(\mathbf{x}_i)$
Define light absorption coefficient γ
while $(t <$ *MaxGeneration)*
for $i = 1 : n$ *all* n *fireflies*
  for $j = 1 : i$ *all* n *fireflies*
    if $(I_j > I_i)$, *Move firefly* i *towards* j *in d-dimension;* **end if**
    Attractiveness varies with distance r *via* $\exp[-\gamma r]$
    Evaluate new solutions and update light intensity
  end for j
end for i
Rank the fireflies and find the current best
end while
Postprocess results and visualization

Figure 1-Pseudo code of Firefly algorithm.

4. The proposed algorithm

The major motivation towards building a feature selection algorithm is to find a subset of features for better performance accuracy. With the traditional wrapper model like the FA, all Fireflies are initialized with randomly selected features, but in the proposed model, all the Fireflies in the swarm will be initialized in a binary sequence. The major steps in the proposed algorithm as follows: -

4.1 Initialization

This step initiates all the Fireflies in the swarm by a random number in the range of (0, 1). These random numbers represent the position of each Firefly, so there are (41) positions of each firefly and are calculated using Equation 1.

$$X = Rand(0,1) \quad (1)$$

where UB and LB represent the upper bound (1.0) and lower bound (0.0), respectively. The generated positions will be converted into a binary sequence using the sigmoid function as follows: -

$$B_i = \begin{cases} 1, & sigmoid\,(X_i) > U\,(0,1) \\ 0, & otherwise \end{cases}$$

where X_i is the position of a Firefly, the sigmoid (X_i) is $1 / (1 + e^{-Xi})$, and U is the uniform distribution. Bi represents the binary sequence, where 1 implies that the feature will be selected, 0 implies the feature will not be selected. The Fireflies are initialized through these steps. Each Firefly has its own position based on the generated number of each one.

4.2 Fitness Function

The fitness function of the proposed algorithm is to minimize the error rate of the classification performance over the validation set of given training data, as shown in Equation 3 while maximizing the number of non-selected features (irrelevant features). To calculate the fitness function, a classifier should be used. In this case, the Naïve Bayesian Classifier was applied to get the accuracy as shown in the Equation 2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Error = 100 - Accuracy \quad (3)$$

After calculating the error, the intensity of each Firefly is calculated using Equation 4.

$$I(F_i) = \frac{1}{1 + Error^2} \quad (4)$$

4.3 Attractiveness Calculation

The attractiveness β of each Firefly can be defined using Equation 5 [17].

$$\beta(r) = \beta_0 \times e^{-\gamma r^2} \quad (5)$$

where r represents the distance between two Fireflies and can be calculated using Equation 6, and β_0 represents the attractiveness at r = 0 (Initial Case) [17].

$$r_{ij} = |X_i - X_j| \quad (6)$$

where X represents the real values of the position of the Fireflies .The distance is calculated using the hamming distance method, by subtracting each bit of Firefly *i* from Firefly *j*. The distance in this method is represented by the difference between the binary strings of the two Fireflies. This method will improve the Firefly algorithm for working with the binary sequence (features) better than working with the continuous values (positions).

4.4 Position Updating

Each Firefly in the swarm moves towards the brighter Firefly; in other words, Fireflies (*Fi*) are attracted by the brighter Firefly. This step can be called position updating which can be determined using Equation 7[17].

$$X_i = P_i + \beta \times (X_j - X_i) + \alpha \times (Rand - \frac{1}{2}) \quad (7)$$

where P_i in the first part of the equation represents the current position, and the second part contains the attractiveness between the position of F_i and F_j. The third part contains the randomization with α, where $\alpha \in [0,1]$. The randomness parameter is decremented by another constant rate δ, where $\delta \in [0.95, 0.97]$, so that at the final stage of the optimization, α has its minimum value as in Equation 8 [17].

$$\alpha = \alpha \times \delta \quad (8)$$

5. Experimental Results

The experiment is divided into two parts; the first part showed the results of the proposed algorithm over different types of testing, while the second part showed the comparison between the proposed algorithm and another two well-known algorithms. Accuracy was measured to analyze the performance of the feature subsets generated by the proposed selection algorithm.

5.1 The results of the proposed algorithm

This part showed the results of the proposed algorithm in terms of accuracy and the number of selected features. The experiments contain two main factors the number of iterations (IT) and number of Fireflies in the swarm or swarm size (SS). The results presented in Tables 2, 3, and 4 showed that the FA can improve the performance of the intrusion detection. Each table contains different values of swarm size (SS) and fixed values of the number of iterations (IT).

From Tables-(2, 3, 4), one can be noticed that the major improvements in the accuracy occurred when the swarm size was increased at the same time the number of iterations affected the results, but with a little improvement.

Table 2-The results of the proposed FA with 250 iterations

SS	*IT*	*Case*	*SF*	*RF*	*No. SF*	*ACC*	*AVE*
10	250	Best	12	29	2,7,8,9,11,21,27,30,35,36,38,40	96.26	95.08
		Worst	15	26	1,2,11,12,13,14,16,19,20,23,25,27,28,35	94.10	
20	250	Best	15	26	1,2,5,6,7,10,11,16,21,27,29,36,38,39,40	96.40	95.53
		Worst	18	23	0,1,2,4,5,16,23,25,27,31,34,35,36,40	94.90	
30	250	Best	13	28	0,1,2,7,8,11,13,19,21,28,38,39,40	96.13	95.64
		Worst	12	29	1,2,3,6,10,11,12,21,27,31,33,36	95.00	
40	250	Best	9	32	2,4,7,11,15,20,38,39,40	96.30	95.67
		Worst	19	22	1,2,3,5,9,10,11,13,18,20,22,25,30,34,35,36,38,39	94.56	

Table 3-The results of the proposed FA with 500 iterations

SS	*IT*	*Case*	*SF*	*RF*	*No. SF*	*ACC*	*AVE*
10	500	Best	16	25	1,2,4,5,7,11,12,15,20,26,30,31,35,38,39,40	96.60	95.31
		Worst	16	25	1,2,5,6,11,12,14,17,20,21,22,25,30,31,32,36	93.90	
20	500	Best	11	30	1,2,7,11,12,20,21,22,38,39,40	96.50	95.56
		Worst	15	26	0,1,2,8,11,12,13,14,19,20,25,26,28,36,39	94.90	
30	500	Best	13	28	1,2,3,5,6,7,10,11,18,19,30,34,36	96.26	95.67
		Worst	17	24	1,2,5,6,9,11,14,15,16,20,23,26,30,33,35,36,37	95.30	
40	500	Best	15	26	1,2,5,7,9,11,12,20,25,26,31,35,36,39,40	96.23	95.70
		Worst	17	24	0,1,2,4,7,8,11,12,13,15,27,31,32,35,36,37,40	95.30	

Table 4-The results of the proposed FA with 1000 iterations

SS	*IT*	*Case*	*SF*	*RF*	*No. SF*	*ACC*	*AVE*
10	1000	Best	14	27	1,2,3,5,7,8,9,11,16,23,28,30,35,36	96.33	95.57
		Worst	18	23	1,2,5,8,9,11,19,20,21,23,27,30,31,33,36,38,39,40	94.90	
20	1000	Best	9	32	2,5,11,14,19,21,26,27,38	96.43	95.78
		Worst	17	24	1,2,3,5,6,8,11,13,15,16,17,18,21,24,27,33,38	95.13	
30	1000	Best	14	27	1,2,8,9,10,18,21,26,27,32,33,37,38,39	96.30	95.87
		Worst	16	25	1,2,5,6,11,12,15,18,25,26,31,35,36,39,40	95.50	
40	1000	Best	14	27	1,2,7,8,11,20,23,24,26,31,35,36,39,40	96.60	96.05
		Worst	14	27	1,2,3,5,11,12,13,18,23,29,35,36,39,40	95.76	

Table-5 summarized the results by comparing the best results of each experiment with the original accuracy (all features).

Table 5-Results of comparing the best results of each experiment with the original accuracy

SS	*IT*	*SF*	*RF*	*ACC*
Original	-	41	0	89.6
40	250	9	32	95.67
40	500	15	26	95.70
40	1000	15	26	96.63

Table-5 showed that the best results were obtained by the maximum swarm size of 40. The results were increasing but with no major difference, when compared with the results based on the number of iterations. We can conclude that the proposed method needed for 40 Fireflies in the swarm but with 250 iterations to decrease the time.

5.2 Benchmarking the proposed method with other algorithms

The proposed IDS model is anomaly based and has two main stages - the pre-processing stage, which involved the wrapper feature selection process that combines BBAL with the detection classifier (NBC); the second stage is the detection step which showed the performance measures obtained by the classifier with previously selected feature subsets. To test the proposed intrusion detection, a personal computer with a core i7 processor, speed 2.2 GHz, and 4 GB of RAM running under windows 10 operating system was used. Also, for the ranking, the proposed algorithm was benchmarked with two other algorithms (Binary Particle Swarm Optimization (BPSO) and Binary Bat algorithm (BBA) [18].

The three algorithms had their individual parameters and use specific values, as follows:
Swarm size = 10, Maximum number of iteration = 200. Table- 6 shows the comparison results of all the algorithms.

Table 6-The results of all the algorithms

Algorithm	*Acc. Rate*	*No. Features*
BPSO [18]	90.63%	22
BBAL [18]	91.61%	15
Proposed Model (BFA + NB)	**94.83 %**	**15**
NB	89.9%	41 (ALL)

6. Conclusion

A wrapper feature selection method was proposed for intrusion detection system. Furthermore, Naïve Bayesian Classifier was used to judge the performance of the proposed method The NSL-KDD dataset was used. The results proved that the movement and randomization of the Firefly algorithm were enhanced by distance calculation through hamming distance method since the Firefly algorithm was initialized by a binary sequence, unlike the standard Firefly algorithm. This enhancement can offer better results in terms of accuracy. Further works will focus on proposing and testing other modifications for improving the meta-heuristic approaches for feature selection problems in general, and the intrusion detection system in particular.

References

1. Paliwal, S., Gupta, R. **2012.** Denial-of-service, probing & remote to user (R2L) attack detection using genetic algorithm. *International Journal of Computer Applications*, **60**(19): 57–62.
2. Rana, F., Najeeb and Ban N. Dhannoon, **2017**. Classification for Intrusion Detection with Different Feature Selection Methods: A Survey (2014-2016). *International Journal*, May 7, **5**: 305-311.
3. Azad, C., Jha, V.K. **2013.** Data mining in intrusion detection: a comparative study of methods, types and data sets. *International Journal of Information Technology and Computer Science*, **5**(8): 75–90.
4. Parazad, S., Saboori, E., Allahyar, A. **2012**. *Fast feature reduction in intrusion detection datasets.* In MIPRO, Proceedings of the 35th International Convention IEEE, pp. 1023–1029.
5. Kang, S.-H. **2015.** A feature selection algorithm to find optimal feature subsets for detecting DoS attacks. IT Convergence and Security (ICITCS), 5th International Conference on. IEEE, pp. 352–354.
6. Nguyen, H.T., Petrovic, S., Franke, K. **2010.** A comparison of feature-selection methods for intrusion detection. *Computer network security*, pp. 242–255.
7. Chebrolu, S., Abraham, A., Thomas, J.P. **2004.** Hybrid feature selection for modeling intrusion detection system. International Conference on Neural Information Processing. Springer, Berlin, Heidelberg, pp. 1020–1025.
8. Chebrolu, S., Abraham, A., Thomas, J.P. **2005.** Feature deduction and ensemble design of intrusion detection systems. *Computers & security*. **24**(4): 295–307.
9. KDD Cup **1999, 2007**, Available on :http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
10. NSL_KDD data set: http://nsl.cs.unb.ca/NSL-KDD/
11. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A. **2009**. A detailed analysis of the KDD CUP 99 data set. Computational Intelligence for Security and Defense Applications, IEEE, pp. 1-6.
12. Chandrashekar, G., Sahin, F. **2014.** *A survey on feature selection methods. Computers & Electrical Engineering*, **40**(1): 16–28.
13. Kayacik, H.G., Zincir-Heywood, A.N., Heywood, M.I. **2005.** Selecting features for intrusion detection: a feature relevance analysis on KDD 99 intrusion detection datasets. Proceedings of the third annual conference on privacy, security and trust Canada.
14. Olusola, A.A., Oladele, A.S. and Abosede, D.O. **2010.** Analysis of KDD '99 intrusion detection dataset for selection of relevance features. Proceedings of the World Congress on Engineering and Computer Science, **1**: 20-22.
15. Xu, Z., King, I., Lyu, M.R.T. and Jin, R. **2010.** Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural networks.* **21**(7): 1033–1047.
16. Cuyon, I. and Elisseeff, A. **2003.** An introduction to variable and feature selection. *Journal of machine learning research* 3.Mar, :1157–1182.
17. Xin-She Yang, **2009.** *Firefly Algorithms for Multimodal Optimization, Stochastic Algorithms: Foundations and Application.* Springer, Berlin, Heidelberg, pp. 169-178.
18. Adriana-Cristina, Enache, Valentin Sgârciu, and Alina Petrescu-Niţă. **2015.** Intelligent feature selection method rooted in Binary Bat Algorithm for intrusion detection. Applied Computational Intelligence and Informatics (SACI), IEEE 10th Jubilee International Symposium on. IEEE, pp. 517-521.