Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2015) 23: 615 – 622

doi:10.3906/elk-1302-53

A new feature selection model based on ID3 and bees algorithm for intrusion detection system

Adel Sabry EESA[1], **Zeynep ORMAN**[2,*], **Adnan Mohsin Abdulazeez BRIFCANI**[3]
[1]Department of Computer Science, Faculty of Science, Zakho University, Duhok City, Iraq
[2]Department of Computer Engineering, Faculty of Engineering, İstanbul University, Avcılar, İstanbul, Turkey
[3]Department of Information Technology, Duhok Technical Institute, Duhok Polytechnic University, Duhok City, Iraq

Received: 08.02.2013 • **Accepted:** 24.04.2013 • **Published Online:** 23.02.2015 • **Printed:** 20.03.2015

Abstract: Intrusion detection systems (IDSs) have become a necessary component of computers and information security framework. IDSs commonly deal with a large amount of data traffic and these data may contain redundant and unimportant features. Choosing the best quality of features that represent all of the data and exclude the redundant features is a crucial topic in IDSs. In this paper, a new combination approach based on the ID3 algorithm and the bees algorithm (BA) is proposed to select the optimal subset of features for an IDS. The BA is used to generate a subset of features, and the ID3 algorithm is used as a classifier. The proposed model is applied on KDD Cup 99 dataset. The obtained results show that the feature subset generated by the proposed ID3-BA gives a higher accuracy and detection rate with a lower false alarm rate when compared to the results obtained by using all features.

Key words: Intrusion detection system, ID3 algorithm, bees algorithm, feature selection

1. Introduction

Recently, security threads have become a crucial problem to overcome for computer networks. For this reason, intrusion detection systems (IDSs) are considered to be one of the most important issues in computer systems. In the last decade, various IDSs [1,2] have been proposed. These systems can generally be classified into 2 categories as anomaly detection and misuse detection. In the misuse detection-based IDS, attacks are detected by comparing them with a very large databases of attack signatures. It searches for a specific signature that is already stored in the database. On the other hand, the anomaly detection-based IDS detects attacks by observing deviations from the normal behavior of the system. It works by training the system with a set of training data to establish some notion of normality and then use the established profile on real data to flag deviations.

Most of the studies of IDSs use pattern recognition techniques to overcome some problems of IDS development such as extraction of attacks and normal signatures from data and detection of unknown attacks during training. Therefore, the IDS can be considered as a pattern discovery and recognition system. The performance of a pattern recognition system can be affected by many parameters like feature extraction and pattern representation. Feature extraction and selection is a part of the dimension reduction used in many fields such as classification task, data mining, object recognition, etc. [3]. Feature extraction creates new features

*Correspondence: ormanz@istanbul.edu.tr

from functions of the original features, whereas feature selection (FS) is the process of selecting a subset of relevant features to be used in model construction.

The use of FS is due to the data, which may contain many redundant or irrelevant features. Given a feature set of size n, the FS problem is to find a minimal feature subset of size m $(m < n)$ that enables the construction of the best classifier with high accuracy [4]. The aim of FS is to reduce the dimension of the dataset and to recognize the corresponding features while satisfying the predictive accuracy. With this specification, it simplifies the dataset and also reduces redundancy in the selected features. In practical problems, FS should be used to avoid irrelevant and/or noisy features.

In recent years, several methods have been proposed for FS. Basiri et al. [4] proposed ant colony optimization based on the selected features for predicting postsynaptic activity in proteins. Wang et al. [5] discussed the shortcomings of conventional hill-climbing rough set approaches to FS and proposed a new FS approach that uses rough sets and particle swarm optimization (PSO). Zhang et al. [6] proposed a quantum PSO and support vector machine (SVM) based on a network intrusion FS wrapper algorithm, considering the relevance among features, which filter-based FS method fails to deal with. Alomari and Othman [7] proposed a combination of the bees algorithm (BA) and SVM for FS. They used the BA for subset generating and the SVM for the classification process in anomaly detection. Kloft et al. [8] proposed a generalization of the support vector data description (SVDD) that can select the best feature combination. SVDD is described as a semiinfinite linear program that can be solved with standard techniques. Fadaeieslam et al. [9] proposed FS based on decision dependent correlation using the SVM classifier. Suebsing and Hiransakolwong [10] applied Euclidean distance for selecting a subset of features to build a model for the detection of known and unknown patterns. Ahmad et al. [11] presented a mechanism for optimal features subset selection that can overcome the drawback of some techniques such as principal component analysis, genetic algorithms, and multilayer perceptrons. Takkellapati and Prasad [12] proposed information gain and triangle area based on the k-nearest neighbor that are used for selecting features by combining a greedy k-means clustering algorithm and SVM classifier to detect network attacks.

In this paper, a new combinational model based on ID3 and BA, ID3-BA, is proposed to select appropriate features for network intrusion detection system. The BA is used for FS while the ID3 algorithm is used as a classifier. The proposed model is tested on KDD Cup 1999 datasets and the obtained results show that the model is capable of minimizing the number of selected features and maximizing the accuracy and the detection rate with a lower false alarm rate (FAR).

The remainder of this paper is organized as follows: in Section 2, a brief description of the ID3 and BA algorithms is given. Section 3 describes the proposed feature selection approach. Section 4 presents the experimental setup and the obtained results. Finally, Section 5 provides the conclusion and future work.

2. ID3 and BA

2.1. The interactive dichotomizer 3 (ID3) algorithm

ID3 is a basic algorithm developed by Quinlan in 1983 [13] that is used to build the classification rules in the form of a decision tree (DT). The resulting tree is constructed top-down from a fixed set of examples. The leaf nodes contain the class name whereas a nonleaf node is a decision node. At each decision node, each attribute is tested to decide how good it classifies the examples. The appropriate attribute is then chosen, while the remaining examples are partitioned by it [14]. Metric information gain is introduced to select the appropriate

attributes for the classification process. By partitioning the examples related to this attribute, entropy reduction is assured, which is also known as the information gain.

2.2. Bees algorithm (BA)

The BA is a swarm-based optimization algorithm developed in 2005 [15]. It mimics the behavior of honeybees to find the best food location. In its basic version, the scout bees look for food locations where they can produce favorable honey. They then carry out localized and organized searches until they find the most efficient possible location for food recovery process. The algorithm can be used for both combinatorial optimization [16,17] and functional optimization [18]. Figure 1 illustrates the pseudocode for a simple BA.

1. Initialize population with random solutions.
2. Evaluate fitness of the population.
3. While (stopping criterion not met) //Forming new population.
4. Select elite bees.
5. Select sites for neighborhood search.
6. Recruit bees for selected sites (more bees for best e sites) and evaluate fitness.
7. Select the fittest bee from each patch.
8. Assign remaining bees to search randomly and evaluate their fitness.
9. End While.

Figure 1. Pseudocode of the BA.

3. The proposed feature selection approach

In this paper, we propose a combination of the BA and ID3 algorithms as a FS approach for a network intrusion detection system. The general principle of the proposed model is shown in Figure 2. The BA is used for generating a subset of features, while the ID3 algorithm is used as a classifier. In order to classify the normal data and the attacked data by using the proposed approach, the following issues must be considered.

3.1. Data representation

The training and the testing data contain both discrete and continuous attributes, and because the standard ID3 algorithm deals only with discrete sets of values, the continuous attributes must be converted to discrete values. This can be done by partitioning the continuous attribute values into a discrete set of intervals.

We used the classical partition method [19] to partition continuous values into 2 intervals as follows: first, the maximum and minimum values for each continuous attribute are determined. The attribute domain is then partitioned into 2 parts, depending on the maximum and minimum values such as $part\_1 \leq (min + (max\text{-}min)/2)$ and $part\_2 > (min + (max\text{-}min)/2)$. In this way, all continuous values will be converted to discrete values.

3.2. Neighborhood search

The KDD data connection records contain 41 features. We ranked these features based on their locations, so we get $ranked\_array = \{1, 2, 3, \ldots 41\}$. Each bee in the population will select a subset of features randomly, and then we get 2 subsets of features as *selected_features* and *unselected_features* where $selected\_features \subset ranked\_\ array$, $unselected\_features \subset ranked\_\ array$, and $selected\_features \cap unselected\_features = \emptyset$. A single

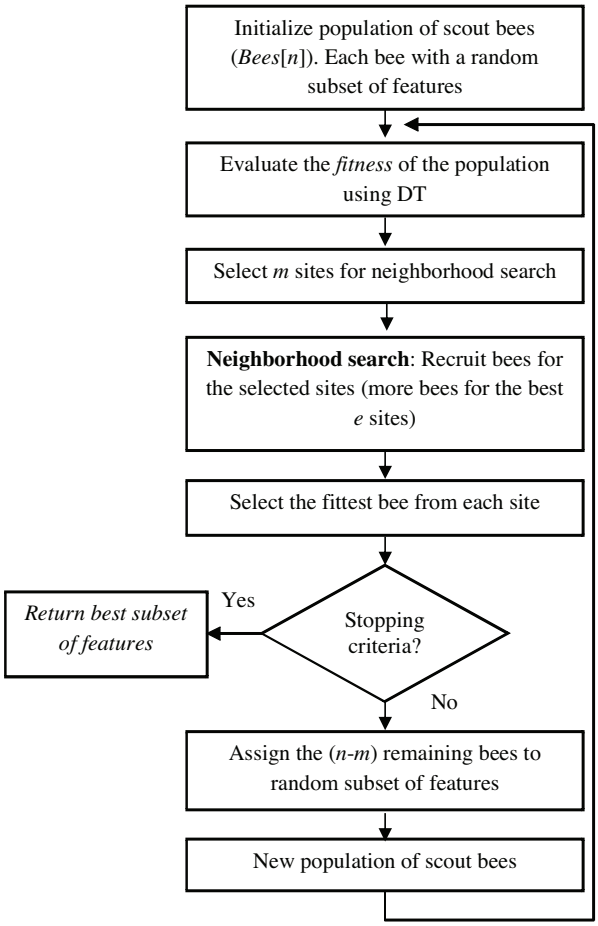

Figure 2. The general principle of the proposed ID3-BA.

random exchange operator is considered for the purpose of neighborhood search. To illustrate, consider a bee with *selected_features* = {1, 2, 3, 4, 5} and *unselected_features* = {6, 7, 8, ..., 41} . A single feature will be chosen randomly from both *selected_features* and *unselected_features* and then a exchange operator will be applied between them. Figure 3 describes the functioning of this operator.

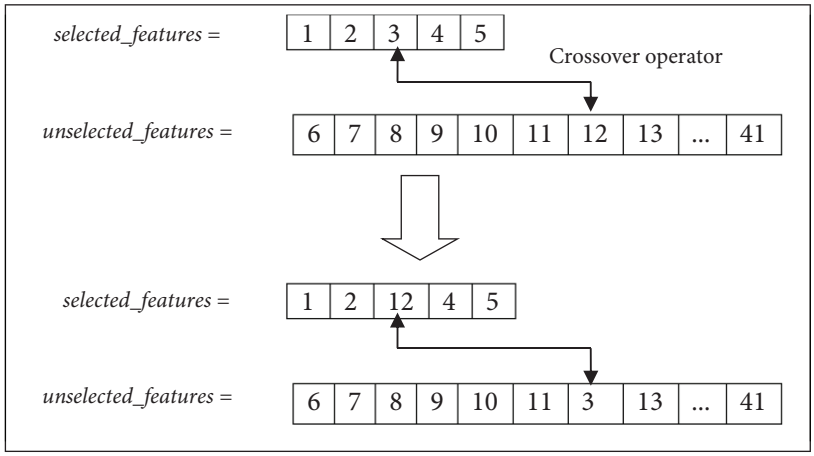

Figure 3. Crossover operator.

4. Experiments and results

4.1. Data source

As mentioned previously, KDD Cup 99 [19,20] is used to evaluate the proposed FS model for intrusion detection. This dataset is a common standard for evaluation of intrusion detection techniques. Ten percent of the KDD Cup 99 training dataset and testing dataset contain about 494,020 and 311,028 connection records, respectively. This amount of data is too large to use in such studies. For this reason, a subset of 10% of the KDD Cup 99 training and testing datasets is extracted, and to keep the proportion of attacks in both the train and test datasets, each attack is divided by 100. For example, the number of IP sweep attacks in the original training and testing data is (1247, 306), while the number of these attacks in the extracted data is equal to (12, 3). Table 1 describes different attack types and their corresponding occurrence number in the training and test data. The number of training data is 4947 and the number of test data is 3117, which are selected randomly. Table 2 describes the number of attacks in the original 10% KDD Cup 99. From Table 1, 'Probing' (41, 42) means that the number of records in the training dataset of probe attacks is equal to 41 connection records, while the number of records in the testing dataset for this attack is equal to 42 connection records.

Table 1. Different attack types and their corresponding occurrence number respectively in the extracted train and test dataset.

Normal (973; 606)			
Probing (41; 42)	**DoS (3915; 2299)**	**U2R (5; 10)**	**R2L (13; 160)**
ipsweep(12;3), Mscan(0;11), Nmap(2;1) Portsweep(11;4) Saint(0;7), Satan(16;16).	apache2(0;8), back(22;11), land(0; 0), mailbomb(0;50), Neptune(1072;580), processtable(0;8), Pod(3;1), udpstorm(0;0), Smurf(2808;1641), Teardrop(10;0).	buffer_overflow(3;1), httptunnel(0;3), loadmodule(0;0), perl(0;0), rootkit(2;2), xterm(0;2), Ps(0;2), Sqlattack(0;0).	ftp_write(0;0), imap(0;0), guesspasswd(2;44), named(0;0), multihop(0;0), phf(0;0), sendmail(0;0), snmpgetattack(0;77), snmpguess(0;24), spy(0;0,) warezclient(10;0), worm(0;0), warezmaster(1;15), xsnoop(0;0), xlock(0;0).

4.2. Evaluation criteria

Three performance measures [7] are used to evaluate the proposed approach: detection rate (DR), false alarm rate (FAR), and accuracy rate (AR). These performance measures are defined with the following equations.

$$DR = \frac{\text{No.of attacks correctly classified as attack}}{\text{Total no. of attacks in the dataset}}$$

$$FAR = \frac{\text{No. of normal events classified as attack}}{\text{Total no. of normal events in the dataset}}$$

$$AR = \frac{\text{No. of correctly classified instances}}{\text{Total no. of instances in the dataset}}$$

Higher values of DR and AR and lower values of FAR show better classification for the IDS.

Table 2. Different attack types and their corresponding occurrence numbers respectively in the original 10% KDD Cup 99 train and test datasets.

Normal (97,277; 60,593)	
Probing (4107; 4166)	**DoS (391,458; 229,853)**
ipsweep(1, 247; 306), mscan(0; 1, 053), nmap(231; 84), portsweep(1, 040; 364), saint(0; 736), satan(1, 589; 1, 633).	apache2(0; 794), back(2, 203; 1.098), land(21; 9), mailbomb(0; 5, 000), neptune(107, 201; 58, 001), pod(264; 87), processtable(0; 759), smurf(280, 790; 164, 091), teardrop(979; 12), udpstorm(0; 2).
U2R(52; 228)	**R2L(1126; 16,189)**
buffer overflow(30, 22), httptunnel(0; 158), guess passwd(53; 4, 367), loadmodule(9; 2), perl(3; 2), perl(3; 2), ps(0; 16), rootkit(10; 13), sqlattack(0; 2), xterm(0; 13).	ftp write(8; 3), imap(12; 1), multihop(7; 18), named(0; 17), phf(4; 2), sendmail(0; 17), snmpgetattack(0; 7, 741), snmpguess(0; 2, 406), spy(2; 0), warezclient(1, 020; 0), warezmaster(20; 1, 602), worm(0; 2), xlock(0; 9), xsnoop(0; 4).
Total train dataset = 494,020 Total test dataset = 311,028	

4.3. Fitness function

The ID3 algorithm is used as a classifier for validating the feature subsets as we mentioned before. Each bee holds a subset of features and it is supported by the ID3 classifier to decide the quality of its features subset. The gauged quality is based on the fitness value, which is defined in Eq. (4):

$$Fitness = \alpha * DR + \beta * (1 - FAR), \tag{4}$$

where $\alpha \in [0, 1]$ and $\beta = 1 - \alpha$ are 2 parameters that show the importance of DR quality and FAR, respectively.

Eq. (4) clearly indicates that DR and FAR qualities have different significance. In our experiment we propose that DR quality is more important than FAR and we set $\alpha = 0.7$ and $\beta = 0.3$.

4.4. Results

The parameters of BA are assumed as follows through the experiments: $n = 40$, $m = 3$, $e = 1$, $nsp = 5$, and $nep = 10$. The simulations have been carried out using C# on a Pentium Dual-Core CPU 2.20 GHz laptop, 2 GB RAM.

Figures 4 and 5 and Table 3 show the results of the ID3-BA approach in terms of DR, FAR, and AR for 10 independent runs. The obtained results show that when the number of features is less than 30, the proposed model gives a higher performance of DR and AR. In all cases, the model gives a lower FAR when compared with the results obtained using all 41 features. Figure 6 describes the ROC curve in terms of the DR and FAR of the intrusion detection based on the proposed ID3-BA.

Table 3. Results of the proposed ID3-BA.

No. of features	FAR	DR	Accuracy rate	Fitness
41	17.685%	71.087%	73.267%	74.455%
35	1.52%	69.183%	74.872%	77.972%
30	1.157%	69.454%	75.16%	78.27%
25	7.421%	84.392%	85.982%	86.848%
20	6.363%	90.987%	91.5%	91.782%
15	3.157%	91.565%	92.59%	93.148%
10	3.702%	91.792%	92.666%	93.143%
5	3.917%	91.02%	92.002%	92.538%

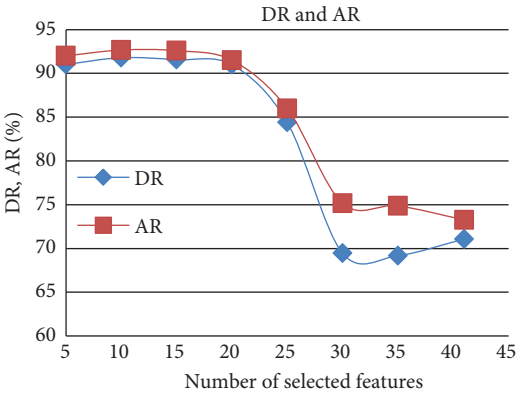

Figure 4. DR and AR per number of selected features using proposed ID3-BA.

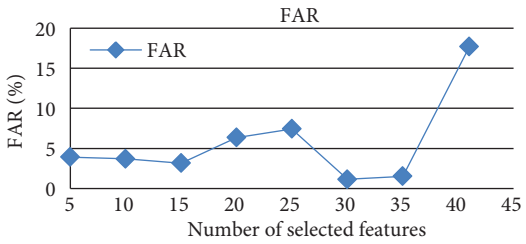

Figure 5. FAR per number of selected features using ID3-BA.

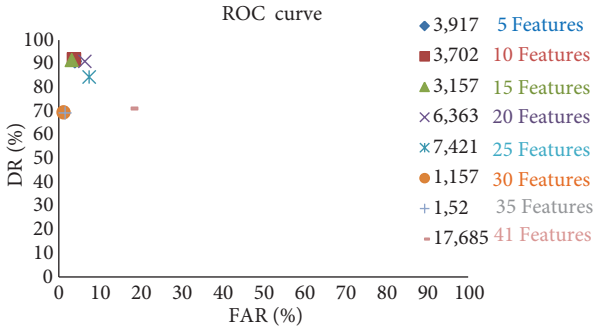

Figure 6. ROC curve of the intrusion detection-based proposed feature selection model.

5. Conclusion and future work

In this paper a novel combination method based on ID3 and BA for subset feature selection is presented. The BA is used for FS while the ID3 algorithm is used as a classifier. Based on the KDD Cup 99 dataset used for the experiments, the obtained results demonstrate that the feature subset produced by the proposed ID3-BA is superior in terms of classification of DR and AR and has a lower FAR when compared with the results obtained using all features. As future work, we will investigate a new approach for the neighborhood search by controlling the number of bees that are responsible for searching the most efficient possible location for food recovery.

References

[1] Betanzos AA, Marono NS, Fortes FMC, Romero JS, Sanchez BP. Classification of computer intrusions using functional networks. a comparative study. In: European Symposium on Artificial Neural Networks; 25–27 April 2007; Bruges, Belgium. pp. 579–584.

[2] Zainal A, Maarof MA, Shamsuddin SM, Abraham A. Ensemble of one-class classifiers for network intrusion detection system. In: Fourth International Conference on Information Assurance and Security; 8–10 September 2008; Washington, DC, USA: IEEE. pp. 180–185.

[3] Al-Ani A. An ant colony optimization based approach for feature selection. In: International Conference on Machine Learning and Cybernetics; 19–21 December 2005; Cairo, Egypt. pp. 3871–3875.

[4] Basiri ME, Ghasem-Aghaee N, Aghdam MH. Using ant colony optimization-based selected features for predicting post-synaptic activity in proteins. In: European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics; 26–28 March 2008; Napoli, Italy. pp. 12–23.

[5] Wang X, Yang J, Teng X, Xia W, Jensen R. Feature selection based on rough sets and particle swarm optimization. Pattern Recogn Lett 2007; 28: 459–471.

[6] Zhang H, Gao H, Wang X. Quantum particle swarm optimization based network intrusion feature selection and detection. In: International Federation of Automatic Control World Congress; July 2008; South Korea. pp. 12312–12317.

[7] Alomari O, Othman ZA. Bees algorithm for feature selection in network anomaly detection. Journal of Applied Sciences Research 2012; 8: 1748–1756.

[8] Kloft M, Brefeld U, Dussel P, Gehl C, Laskov P. Automatic feature selection for anomaly detection. In: First ACM Workshop on AISec; 27–31 October 2008; Alexandria, VA, USA. pp. 71–76.

[9] Fadaeieslam MJ, Minaei-Bidgoli B, Fathy M, Soryani M. Comparison of two feature selection methods in intrusion detection systems. In: International Conference on Computer and Information Technology; 16–19 October 2007; Fukishima, Japan: IEEE. pp. 83–86.

[10] Suebsing A, Hiransakolwong N. Euclidean-based feature selection for network intrusion detection. In: International Conference on Machine Learning and Computing; 26–28 February 2011; Singapore. pp. 222–229.

[11] Ahmad I, Abdulah AB, Alghamdi AS, Alnfajan K, Hussain M. Feature subset selection for network intrusion detection mechanism using genetic eigen vectors. In: International Conference on Telecommunication Technology and Applications; 2–4 May 2011; Sydney, Australia. pp. 75–79.

[12] Takkellapati VS, Prasad GVSNRV. Network intrusion detection system based on feature selection and triangle area support vector machine. International Journal of Engineering Trends and Technology 2012; 3: 466–470.

[13] Salzberg SL. Book review: C4.5: Programs for Machine Learning, by J. Ross Quinlan, Morgan Kaufmann Publishers, 1993. Mach Learn 1994, 16: 235–240.

[14] Lindell Y, Pinkas B. Secure multiparty computation for privacy-preserving data mining. The Journal of Privacy and Confidentiality 2009; 1: 59–98.

[15] Pham DT, Ghanbarzadeh A, Koc E, Otri S, Rahim S, Zaidi M. The Bees Algorithm. Technical Note. Cardiff, UK: Manufacturing Engineering Centre, Cardiff University, 2005.

[16] Pham DT, Afify A, Koc E. Manufacturing cell formation using the bees algorithm. In: Innovative Production Machines and Systems Virtual Conference; 2–13 July 2007; Cardiff, UK.

[17] Pham DT, Koc E, Lee JY, Phrueksanant J. Using the bees algorithm to schedule jobs for a machine. In: International Conference on Laser Metrology, CMM and Machine Tool Performance; 25–28 June 2007; Cardiff, UK. pp. 430–439.

[18] Pham DT, Ghanbarzadeh A, Koç E, Otri S, Rahim S, Zaidi M. the bees algorithm – a novel tool for complex optimization problems. In: Innovative Production Machines and Systems Virtual Conference; 2006. pp. 454–459.

[19] Sabry A. A comparative study among several modified intrusion detection system techniques. MSc, Duhok University, 2009.

[20] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: IEEE Symposium on Computational Intelligence in Security and Defense Applications; 8–10 July 2009; Ottawa, Canada. pp. 1–6.