Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 5, No. 4, November 2014), Pages: 43-52
www.jacr.iausari.ac.ir

Intrusion Detection System in Computer Network Using Hybrid Algorithms (SVM and ABC)

Bahareh Gholipour Goodarzi✉[1], Hamid Jazayeri[2], Soheil Fateri[1]

1) Computer Engineering Department, Islamic Azad University, Babol Branch, Babol, Iran
2)Electrical and Computer Engineering Department, Nushirvani University of Technology, Babol, Iran

gholipour.bahar@gmail.com; jhamid@nit.ac.ir; fateri@gmail.com

Received: 2014/02/12; Accepted: 2014/03/15

Abstract

In recent years, the needs of the Internet are felt in lives of all people. Accordingly, many studies have been done on security in virtual environment. Old technics such as firewalls, authentication and encryption could not provide Internet security completely; So, Intrusion detection system is created as a new solution and a defense wall in cyber environment. Many studies were performed on different algorithms but the results show that using machine learning technics and swarm intelligence are very effective to reduce processing time and increase accuracy as well. In this paper, hybrid SVM and ABC algorithms has been suggested to select features to enhance network intrusion detection and increase the accuracy of results. In this research, data analysis was undertaken using KDDcup99. Such that best features are selected by Support vector machine, then selected features are replaced in the appropriate category based on artificial bee colony algorithm to reduce the search time, increase the amount of learning and improve the authenticity of intrusion detection. The results show that the proposed algorithm can detect intruders accurately on network up to 99.71%.

Keywords: *Intrusion Detection System, Support Vector Machine, Classification, Bee colony Algorithm*

1. Introduction

In the last decade, advances in technology allowed computers to be remotely managed and also provided a gateway to all kind of information through the Internet. In addition, organizations face the problem of keeping their information protected, available and reliable. Besides, obtaining confidential and privileged information coupled with the challenge to get them became even more exciting to those people interested in obtaining unauthorized access in computer networks. Because of convenience of internet, it is easy to get access to attack knowledge and methods. At present, hackers are unnecessary to have a wide knowledge of specialized knowledge, and annual internet attack cases are increasing to a great extent. So solution felt for organization and encounter to problems such as data protection and data accuracy [1], [2]. Traditional protection techniques such as user authentication, data encryption, avoiding programming errors and firewalls are used as the first line of defense forcomputer security but none of them are able to protect network completely [3], [4]. Therefore the concept of intrusion detection system was first suggested in a technical report by

Anderson in 1980. Intrusion detection system is to supervise and control all cases happening to computer system or network system, analyze any signal arising from related safety problems, send alarms when safety problems occur, and inform related personnel or units to take relevant measures to reduce the possible risks. Its framework includes three parts:

1. Information collection: Data collection: the source of these collected data can be separated into host, network and application, according to the position.
2. Analysis engine: Analysis engine is able to analyze whether or not there are symptom of any intrusion.
3. Response: Take actions after analysis, record analysis results, send real-time alarm, or adjust intrusion detection system [1].

Three types of data are used by IDSs. These are network traffic data, system level test data and system status files. IDSs are used in order to stop attacks, recover from them with the minimum loss or analyze the security problems so that they are not repeated [3].

Intrusion Detection Systems are divided in to two general categories for to detect attackers: 1. According to different data source, 2. According to different analysis method. The first category is again divided to host based IDS and the Network based IDS and second category also includes misuse detection or signature based detection and anomaly detection.

Data in host based IDS comes from record of various activities of host, including audit record of operating system, system log, application programs information and so on. Data in network based IDS is mainly collected network generic stream going through network segments, such as Internet packets. The comparison between these two methods can note that host based IDS judge whether or not the host is intruded more accurately, detect attackers under encrypted network environment, does not need additional hardware as advantage and it may affect system efficiency of monitored hosts, higher cost as disadvantage. For network based IDS can be noted to low cost, it can detect attacks that cannot be done by host based IDS as advantage and the flux is large, and some packets may be lost, and it cannot detect all packets in network, In large-scale network, it requires more rapid CPU and more memory space to analyze bulk data. It cannot deal with encrypted packets, and it may not receive attack information in encrypted packets they are accordingly its disadvantage.

Misuse Detection can transform the information of attack symptom or policy disobeying into state transition-based signature or rule, and such information is stored in signature database. To judge whether or not it is attack, pre-treated case data should be first compared with the signature of signature database, and those conforming to attack signature data can be judged as attack. High detection rate and low false alarm rate for known attacks are its advantage; however, its detection capacity is low for un-known detection methods, and attack database should be renewed on a regular basis.

Anomaly Detection may establish a profile for normal behavior of users, which comes from statistics data of users in the former period; when detection is performed, the profiles are compared with actual users' data, if the offset is below threshold value, user's behavior can be considered normal, and it has no intention of attacks; if the offset is above threshold value, user's behavior can be considered abnormal. Detection rate of the method is high, and it is more likely to detect un-known attacks, but misjudgment rate is also high [1], [3], [5]. Many researchers have proposed and implemented various

models for IDS but they often generate too many false alerts due to their simplistic analysis. An attack generally falls into one of four categories:

- DoS attack: Denial of Service attack results by preventing legitimate requests to a network resource by consuming the bandwidth or by overloading computational resources.
- Probing attack: this is a type of attack which collect information of target system prior to initiating an attack.
- User to Root (U2R) attack: In this case, an attacker starts out with access to a normal user account on the system and is able to exploit the system vulnerabilities to gain root access to the system.
- Root to Local (R2L) attack: In this, an attacker who doesn't have an account on a remote machine sends packet to that machine over a network and exploits some vulnerabilities to gain local access as a user of that machine [4], [5], [6].

In order to solve the problems mentioned above, numbers of anomaly detection systems are developed based on machine learning techniques. These systems use a "normal behavior" to detect those unexpected attacks [2], [7], [8]. However, it seems that none of them is able to detect all kind of intrusion attempts efficiently in terms of detection rate and false alarm rate. Hence, the need is to combine different classifiers as a hybrid data mining strategy to enhance the detection accuracy of the model built in order to make efficient intelligent decisions in identifying the intrusions [6]. This research used SVM and swarm intelligence algorithms. Last results show that combination of SVM and swarm Intelligence always makes IDS more consistent, stronger and faster which has good results in most classes of attack. Rest of the paper is organized as follows: Section 2 and 3 give overview of Support Vector Machine and Artificial Bee Colony respectively. Section 4 describes KDDcup99 Data set. Section 5 presents Implementation of proposed algorithm and results. Finally the paper is concluded with their future work in section 6.

2. Support Vector Machine

Support vector machine is proposed by Vapnik in 1995 [1]. In recent years Investigations shown that this method is one of the strongest and more accurate methods in machine learning algorithms that is used for classification, regression and prediction. High efficiency and appropriate generalization of SVM led to this method found great popularity among researchers [1], [9], [10]. Performance of this method is that SVM first maps the input vector into a higher dimensional feature space and then obtain the optimal separating hyper-plane in the higher dimensional feature space.
Moreover, a decision boundary, i.e. the separating hyper-plane, is determined by support vectors rather than the whole training samples and thus is extremely robust to outliers. It should be noted that support vectors are the training samples close to decided boundary. SVM is able to generate user-defined parameters, which is called penalty factor. It allows users to make a tradeoff between the number of misclassified samples and the width of a decision boundary. SVM is able to minimize the structural risks of statistical learning theory. Therefore it shows the ability of good learning and skills of generalization of two features in intrusion detection that indicate a high dimensional datasets or altered to the top understanding [8], [11].

3. Bee Colony Algorithm

The term Swarm Intelligence (SI) was first introduced by Beni in the context of cellular robotics system. Methodologies, techniques and algorithms that this research field embraces draw their inspiration from the behavior of insects, birds and fishes, and their unique ability to solve complex tasks in the form of swarms, although the same thing would seem impossible in individual level. Indeed, single ants, bees or even birds and fishes appear to have very limited intelligence as individuals, but when they socially interact with each other and with their environment, they seem to be able to accomplish hard tasks such as finding the shortest path to a food source, organizing their nest, synchronize their movement and travel as a single coherent entity with high speed etc. This achievement becomes even more centralized authority (e.g., the queen of the hive) dictating any of this behavior. Applications of this can be found in NP-hard optimizations problems such as the traveling salesman, the quadratic assignment, scheduling, vehicle routing etc. Also, it is a common ground that intrusion detection problems in general and anomaly detection IDS in particular have to cope with huge volume and high dimensional datasets, the need for real time detection, and with diverse and constantly changing behavior. This is where computation intelligence comes into play and converges with the IDS realm [11].

In recent years, several biological and natural processes had been influencing the methodologies in science and technologies. One of them is the newly developed swarm intelligence algorithm based on the behavior of the bees namely Artificial Bee Colony (ABC) [12], [13]. Artificial Bee Colony is proposed in 2005 by Karaboga. This algorithm like other meta-heuristic algorithms was introduced to handle unconstrained benchmark optimization function. Then extended version of the ABC algorithm was presented to handle constrained optimization problems [14]. Artificial Bee Colony contains three groups of bees: employed, onlooker and scout bees. Employed bees seek food situation (solution) randomly then employed bees share their food source information with onlooker bees waiting in the hive by dancing on the dancing area. Duration of dancing depends on amount of nectar (fitness function value) which is exploited by employed bees. Onlooker bees watch different dances before selecting a food source to find the best food source. In the ABC algorithm, onlookers and employed bees perform the exploration process in the search space, while, on the other hand, scouts control the exploration process. Four characteristics about behavior of honey bees can be expressed as follows:

- **Positive feedback:** As nectar amount of food sources increases, number of onlookers visiting them increases.
- **Negative feedback:** The exploitation process of poor food sources is stopped by waggle dance.
- **Fluctuations:** The scouts carry out a random search process for discovering new food sources.
- **Multiple interactions:** Bees share their information about food sources on the dance area [12], [14], [15].

Implementation of bee algorithm is shown in Figure 1.

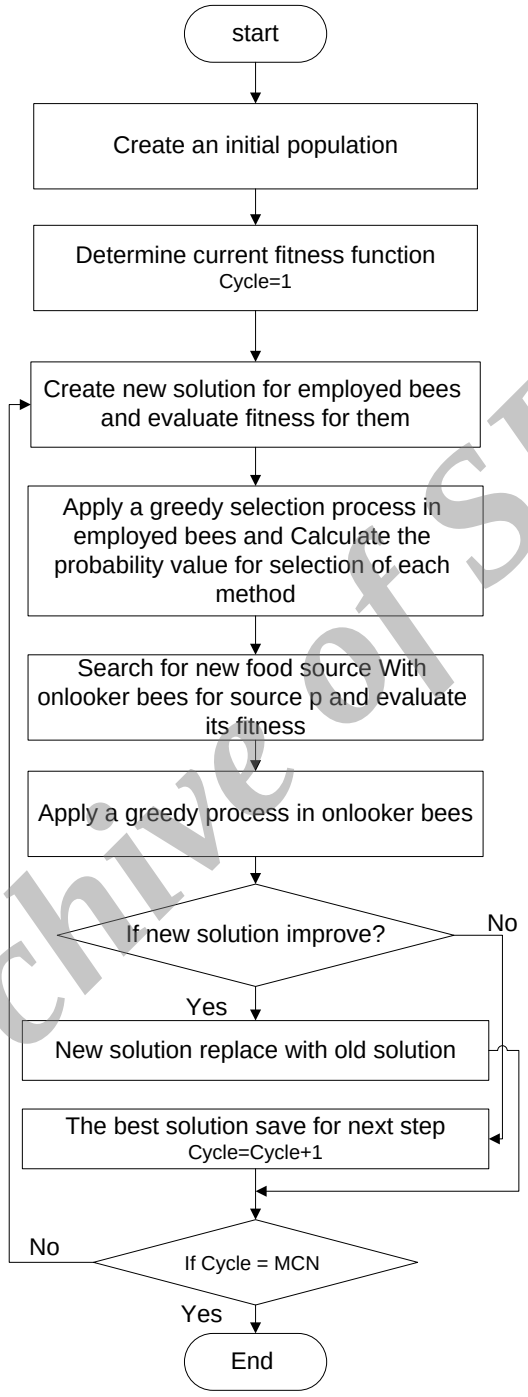

Figure1. Implementation of bee algorithm

4. KDD Cup99 Dataset

The first and most important part of data mining knowledge is collecting correlate data with research field. The best source to detect anomalous behavior and influence in network from normal behavior is the traffic of network. Since 1999, KDD'99 [3] has been the most wildly used data set for the evaluation of anomaly detection methods. This data set is prepared by Stolfo et al. and is built based on the data captured in DARPA'98 IDS evaluation program. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic, which can be processed into about 5 million connection records, each with about 100 bytes. The two weeks oftest data have around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type [1], [5], [16]. These features include continues, discrete and symbolic states and they are grouped into four categories:

- *Basic Features:* Basic features can be derived from packet headers without inspecting the payload.
- *Content Features*: Domain knowledge is used to assess the payload of the original TCP packets. This includes features such as the number of failed login attempts;
- *Time-based Traffic Features*: These features are designed to capture properties that mature over a 2 second temporal window. One example of such a feature would be the number of connections to the same host over the 2 second interval;
- *Host-based Traffic Features*: Utilize a historical window estimated over the number of connections – in this case 100 – instead of time. Host based features are therefore designed to assess attacks, which span intervals longer than 2 seconds [17].

5. Proposed Method Implementation and Results

In this research we used Support Vector Machine and Bee colony algorithms. The algorithm steps are as followed:

- Preprocessing
- Using SVM for feature selection and classification
- Using Bee algorithm for improving last result

Because of unimportant and noisy data in existing datasets, the first and most important step in data mining is data pre-processing until the results have adequate authenticity and accuracy. In this study KDDcup99 dataset, WEKA 3.6.9 and MATLAB 2011 software has been used to select features.

After data pre-processing, for selecting the best feature and classifying them for intrusion detection, LIBSVM function have been investigated. LIBSVM function need discrete data with true value for each attribute. So, in first step, discretize function applied on existing dataset for discretization. Then LIBSVM Cross-Validation method were used on 41 features of KDD dataset with fold = 10. Performance of this method is that nine parts are considered as training and just one part as test. After applying, this

algorithm lead to select 14 features for next step. Table 1 shows the selected features and their number in dataset.

Table 1.Selected features from KDDcup dataset

Feature name	Type	Description
Service	Discrete	• Destination Service
src_bytes	Continuous	• Bytes Sent from Source to Destination
dst_bytes	Continuous	• Bytes Sent from Destination to Source
Land	Discrete	• 1if connection is from/to the same host/port ; 0 otherwise
Wrong_fragment	Continuous	• Number of Wrong Fragment
Num_failed_logins	Continuous	• Number of Failed Login Attempts
Root_shellotherwise	Discrete	• 1 if root shell is obtained; 0 otherwise
Count	Continuous	• Number of connections to the same host as the current connection in the past two seconds
Srv_count	Continuous	• Number of connections to the same service as the current connection in the past two seconds (same service connection)
Srv_rerror_rate	Continuous	• % of connections that have “SYN” errors (same service connections)
Same_srv_rate	Continuous	• % of connections to the same service (same service connections)
Diff_srv_rate	Continuous	• % of connections to different services
Dst_host_same_src_port_rate	Continuous	• % of connections to the current host having the
Dst_host_srv_serror_rate	Continuous	• % of connections to the current host and specified service that have an S0 error

The results of applying this algorithm show that 490853 of 494020 features are in correct classes. It indicates that 99.36% of instances are correctly classified and only 0.64% of instances are classified incorrectly.

Then, in next step, Bee algorithm is applied on results which are obtained in previous step. This algorithm consists of 3 parts: employed bees, onlooker bees and scout bees. Pseudo code of Artificial Bee Colony algorithm is given as below:

1 Initialize the population of solution x_i
2 Evaluate the population by (1)
3 Cycle $\longleftarrow 1$

4 Repeat
5 Produce and evaluate new solution v_i for the employed bees by (2)
6 Apply the greedy selection process for the employed bees
7 Calculate the probability p_i for the solutions v_i refer to (3)
8 Produce the new solutions v_i for the onlookers from the solutions p_i and evaluate them
9 Apply the greedy selection process for the onlookers
10 Determined the abandoned solution for the scout, if exists, and replace it with a new randomly produced solution x_i refer to (2)
11 Memorize the best solution achieved so far
12 Cycle = cycle + 1
13 Until cycle = MCN

By implementing the above code in MATLAB 2011, detection rate and accuracy of the algorithm are achieved according to the following relations. In these relations True Positive (TP): shows the number of attacks which are detected correctly and they had been attacked actually. True Negative (TN): indicates the number of detected normal instances which had been normal in fact. False Positive (FP): or the false alarm rate. It is the number of detected attacks which had been normal. False Negative (FN): the number of detected normal which had been attacks actually. In other words, one can say that these attacks are the intrusion detection systems aim.
Evaluation criteria of detection rate and accuracy based on TP, TN, FP, FN values are defined as follows:

$$\text{Detection Rate} = \frac{TP}{TP+FN} \times 100 \quad (1)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (2)$$

Table 2 shows the results of bee algorithm based on the selected features from previous step. In this table, results are classified according to variety of attacks. We can see that proposed algorithm detection of dos and probe attacks have good results but detection of U2R and R2L attacks are not strong enough.

Table 2.Proposed algorithm results

Attack Type	Detection rate (%)	Accuracy(%)
DoS	99.70	99.69
Probe	96.33	99.66
U2R	56.37	76.00
R2L	64.56	87.37
Normal	99.72	99.70

Figure 2 shows the decrease of error rate in using bee algorithm by 1000 generation. As we can see, bee algorithm moves towards optimal solution by reducing the errors. Besides, Table 3 illustrates a comparison between proposed algorithm and other algorithms results.

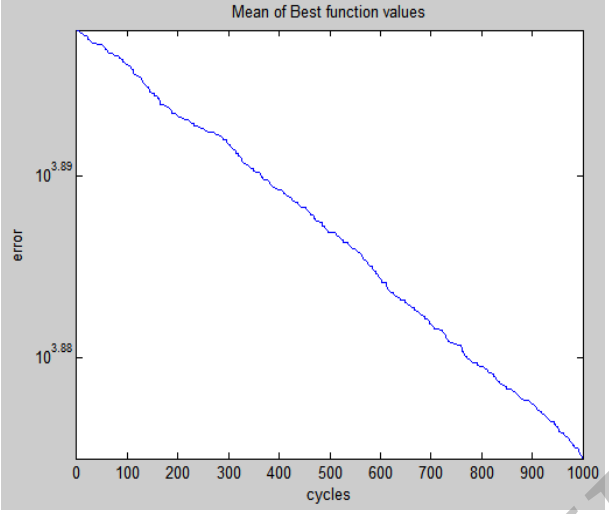

Figure 2.Reducing error based on using Bee algorithm in 1000 generation

Table 3.Comparison between proposed algorithm and other algorithms

Algorithm	Accuracy (%)
SVM-ABC	99.71
SVM-PSO	98.76
SVM-GA	97.35
SA-SVM	99.42
SVM-DT	99.70

Tables 2 and 3 demonstrate the high accuracy of the proposed algorithm compared to other algorithms. They show the proposed algorithm is able to detect attackers correctly to intrude to network amount of 99.71%.

6. Conclusion and Future Works

In recent years, by spread of using the Internet, need of information security has been felt more than ever to prevent personal and confidential information from unauthorized intrusion. The different approaches introduce for intrusion detection. The results of research in different fields show that hybrid methods have better and more accurate results. In this research used hybrid SVM-ABC algorithm. In this approach we have used SVM algorithm for classification features and to improve classification results, Bee algorithm are applied. The proposed algorithm has been performed on the KDDcup99 dataset and the results indicate that our algorithm improves the results compared to other swarm intelligence algorithm in intrusion detection system.
the proposed method have not been very successful in detect R2L and U2R attacks, so future research can pay more attention to these two particular attacks. Moreover, reducing the number of selected features by maintaining the accuracy of the results can be considered as a challenge to future research.

7. References

[1] Wu, S.Y. and E. Yen, *Data mining-based intrusion detectors.* Expert Systems with Applications, 2009. **36**(3): p. 5605-5612.

[2] Pereira, C.R., et al., *An Optimum-Path Forest framework for intrusion detection in computer networks.* Engineering Applications of Artificial Intelligence, 2012.

[3] Aydın, M.A., A.H. Zaim, and K.G. Ceylan, *A hybrid intrusion detection system design for computer network security.* Computers & Electrical Engineering, 2009. **35**(3): p. 517-526.

[4] Peddabachigari, S., et al., *Modeling intrusion detection system using hybrid intelligent systems.* Journal of network and computer applications, 2007. **30**(1): p. 114-132.

[5] Mukherjee, S. and N. Sharma, *Intrusion Detection using Naive Bayes Classifier with Feature Reduction.* Procedia Technology, 2012. **4**: p. 119-128.

[6] Panda, M., A. Abraham, and M.R. Patra, *A Hybrid Intelligent Approach for Network Intrusion Detection.* Procedia Engineering, 2012. **30**: p. 1-9.

[7] Tsai, C.F., et al., *Intrusion detection by machine learning: A review.* Expert Systems with Applications, 2009. **36**(10): p. 11994-12000.

[8] Tsai, C.F. and C.Y. Lin, *A triangle area based nearest neighbors approach to intrusion detection.* Pattern recognition, 2010. **43**(1): p. 222-229.

[9] Tribak, H., et al. *Statistical analysis of different artificial intelligent techniques applied to Intrusion Detection System*. in *Multimedia Computing and Systems (ICMCS), 2012 International Conference on*. 2012. IEEE.

[10] Ektefa, M., et al. *Intrusion detection using data mining techniques*. in *Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on*. 2010. IEEE.

[11] Kolias, C., G. Kambourakis, and M. Maragoudakis, *Swarm intelligence in intrusion detection: A survey*. computers & security, 2011. **30**(8): p. 625-642.

[12] Karaboga, D., et al., *A comprehensive survey: artificial bee colony (ABC) algorithm and applications.* Artificial Intelligence Review, 2012: p. 1-37.

[13] Bae, C., et al., *A NOVEL ANOMALY-NETWORK INTRUSION DETECTION SYSTEM USING ABC ALGORITHMS.* INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING INFORMATION AND CONTROL, 2012. **8**(12): p. 8231-8248.

[14] Abu-Mouti, F.S. and M.E. El-Hawary. *Overview of Artificial Bee Colony (ABC) algorithm and its applications*. in *Systems Conference (SysCon), 2012 IEEE International*. 2012. IEEE.

[15] Wang, J., T. Li, and R. Ren. *A real time idss based on artificial bee colony-support vector machine algorithm*. in *Advanced Computational Intelligence (IWACI), 2010 Third International Workshop on*. 2010. IEEE.

[16] Tavallaee, M., et al. *A detailed analysis of the KDD CUP 99 data set*. in *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009*. 2009.

[17] Kayacik, H.G., A.N. Zincir-Heywood, and M.I. Heywood. *Selecting features for intrusion detection: a feature relevance analysis on KDD 99 intrusion detection datasets*. in *Proceedings of the third annual conference on privacy, security and trust*. 2005. Citeseer.