International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2018.32724

Volume 4, Issue 5
May - 2018

A novel hybrid GA and SVM with PSO feature selection for intrusion detection system

Mehdi Moukhafi[1*], Khalid El Yassini[2] and Seddik Bri[3]

[12]Department of Mathematics and Computer Sciences,

Faculty of Sciences, Moulay Ismail University,

[3]Department of Electrical Engineering,

ESTM, Moulay Ismail University,

Meknes,

Morocco

ABSTRACT

The computer network technologies are evolving fast, and the development of internet technology is more quickly, people more aware of the importance of the network security. Network security is the main issue of computing because the numbers of attacks are continuously increasing. For these reasons, intrusion detection systems (IDSs) have emerged as a group of methods that combats the unauthorized use of a network's resources. Recent advances in information technology, especially in data mining, have produced a wide variety of machine learning methods, which can be integrated into an IDS. This study proposes a new method of intrusion detection that uses support vector machine optimizing by a genetic algorithm. to improve the efficiency of detecting known and unknown attacks, we used a Particle Swarm Optimization algorithm to select the most influential features for learning the classification model.

Key Words: *Machine learning, Intrusion Detection System (IDS), Genetic Algorithm, Support Vector Machine, Particle Swarm Optimization, kdd99.*

1. INTRODUCTION

Due to the tremendous growth of the Internet and Network based services, the severity o f network based computer attacks have significantly increased. Thus, an intrusion detection system (IDS) play a vital role in network security. Intrusion detection system tries to detect computer attacks by examining various data records. the IDS was presented for the first time by Anderson in 1980 [1], and later formalized by Denning [2], it can be used in the global security politics, which includes other protection tools, such as firewalls and anti-virus software; Thus, it is important to take the advantage of these tools collaboration and complementarity. Actual IDS's based on heuristic rules, such as Snort are signature based system. The problem with such a system is that it cannot detect novel attacks whose signature is not available and hence generates a high rate of alerts. where constantly the environments were changing, the major drawback of approaches based signature is that they only detect known attacks, which implies a frequent updating of the rules database and the time for the implement. To overcome the mentioned problem above, many data mining techniques have been developed [3].

The data mining techniques are better applied equally for an anomaly intrusions detection, also for a knowledge-based intrusions detection [4]. The statistical analysis of the normal system behavior is one of the first approaches to intrusion detection. The statistics are used mathematically to describe an observed mechanism. Generally, the observations allow to get a rough description. For this, the value of certain observations is considered random variables. For each of its comments, a statistical model is used to describe the set of the corresponding random variable distributions.

Learning algorithms can play an important role in detecting attacks (known or unknown). Additionally, the IDS's performances is considerably improved at the network level. SVM obtains a good detection performance in terms of classifying the flow of a network into normal or abnormal behaviors. Feng et al. [5] introduced an approach combining SVM with self-organized ant colony network. Kuang et al. [6] propose a solution based on a combination of the SVM model with kernel principal component analysis (KPCA) and genetic algorithm. KPCA was used to reduce the dimensions of feature vectors, whereas GA was employed to optimize the SVM parameters. Al-Yaseen et al. [7] Propose a solution based on hybrid SVM and Extreme Learning Machine model Learned with data set built by a modified K- means The modified K-means is used to build new small training datasets representing the entire original training dataset.

The rest of this work is organized as follows: Section 1 describes the PSO, SVM and GA methods, section 2 proposed architecture, section 3 simulation of results and evaluation of the algorithm. Section 4 conclusion and future work.

2. THE USED METHODS

2.1. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a stochastic optimization method, for the nonlinear functions, inspired by the social behavior of insect colonies, bird flocks, fish schools and other animal societies, PSO was invented by Russell Eberhart and James Kennedy [8] in 1995. Originally, the two began developing software simulations birds flocking around food sources, later after realizing that their algorithm solve optimization problems, they present [9] a discrete binary PSO algorithm developed from the previous PSO and operating in continuous variables.

PSO is an iterative algorithm to find the best solution based on a population composed of many particles. For example, a flock of birds (particles) encircling an area where they can feel a hidden source of food. Whoever the closest to food warn others birds to move toward its direction. If any of the other birds circling closer to the target more than the first, it warbles stronger and the others move towards him. This scheme continues until one of the birds find food.

A particle (candidate solution) that may move to the optimal position by updating its position and its speed. The speed of movement of a particle can be updated by the weight of inertia, cognitive learning factor, and the values of social learning factors.

2.2. Support Vector Machine

Support Vector Machine (SVM) is one of the most popular supervised machine learning algorithms. This is a classification model by evaluating data and identify patterns that retains excellent long generalization capabilities with an integrated resistance to overtraining. This generalization is based on solid theoretical foundations introduced by Vapnik [10]. An SVM model as shown in Figure 1 is an illustration of examples of points in two-dimensional space, where instances of different groups are separated by an area called margin.

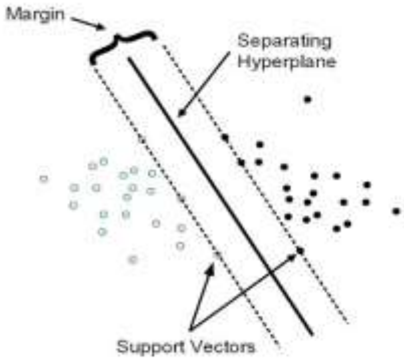

Figure 1: Classical example of SVM linear classifier

In the classification of support vector, the separation function is a linear combination of grains as given in equation (1) and are in contact with the support vector,

$$\mathbf{f(x)} = \sum_{i \in S} \mu_i \, y_i x_i^t x + b \qquad (1)$$

Where μ_i is a Lagrange xi factor is training models, yi {+ 1, -1} is the corresponding class labels and S denotes the set of support vectors.

DOI: 10.31695/IJASRE.2018.32724

2.3. Genetic algorithm

The genetic algorithm (GA), refers to a model introduced and investigated by Holland, J. H. [11] It is a general adaptive optimization search methodology based on a direct analogy to Darwinian's principle of evolution and survival of fittest to optimize a population of candidate solutions towards a predefined fitness.
The The procedure for a genetic algorithm is:

- Initialization - Create an initial population. This population is usually randomly generated of n chromosomes (suitable solutions for the problem)
- Evaluation - Each chromosome x of the population is then evaluated and we calculate a 'fitness' for that individual.
- Selection - Select two parent chromosomes from a population, based on their fitness (the better fitness, the bigger chance to be selected)
- Crossover - create new individuals by combining aspects of our selected parent.
- Mutation - The algorithm creates mutation children by randomly changing the genes of individual parents. Mutation typically works by making very small changes at random to an individual genome.
- Test - If the end condition is satisfied, stop, and return the best solution in current population, else return to selection step.

3. PROPOSED MODEL FOR INTRUSION DETECTION

3.1. PSO features selection

From an artificial intelligence perspective, create a classifier means creating a template for data, or perfect for a model is to be as simple as possible. Reducing the number of parameters, then reduces the number of parameters necessary for the description of this model.

- It improves the classification performance: learning time, his speed and power of generalization.
- It increases the comprehensibility of data. This data selection is to select an optimum subset of relevant variables from a set of original variables.

The KDD99 has 41 variables, it is relatively a large number to be processed by the classifier, the latter in the learning phase cannot complete execution within a reasonable time, then the selection can reduce the feature space. We used PSO to select the optimum features, that have the most impact on the prediction of the model, so we reduce the number of fields from 41 to 16 table 1.
Selected features:

2,3,4,5,6,8,11,14,23,26,29,30,35,36,37,38

Table 1: Comparison of number of features between the data set and subset

Data set	0,1,18,10,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0,0,0,0 ,1,0,0,150,25,0.17,0.03,0.17,0,0,0,0.05,0,normal
Subset	1,18,10,491,0,0,0,0,2,0,1,0,0.03,0.17,0,0,normal

3.2. Architecture of proposed IDS

To implement our proposed approach, the RBF kernel function is used for the SVM classifier because the RBF kernel function has an excellent performance for the management of higher-dimensional data and requires that only two parameters:

- C (penalty parameter): This parameter, common to all SVM kernels, trades off misclassification of training examples against simplicity of the decision surface.
- γ(gamma parameter): It is a specific parameter to RBF kernel function, gamma defines how much influence a single training example has.

The parameters (C and g) used as input attributes must be optimized using genetic algorithm. To precisely establish a GA-SVM based intrusion detection system, the following main steps (as shown in Fig. 2) must be proceeded. The detailed explanation is as follows:

Step 1 : Features selection using PSO for a training data set (kdd99_p)

Step 2 : Initialization of population by a genetic algorithm

Step 3 : Genetic algorithm process: In this step, the system searches for better solutions by genetic operations, including selection, crossover, mutation (Described in the previous session)

Step 4 : Training SVM classifier using optimized parameters (C,γ)

Step 5 : Fitness evaluation. For each chromosome representing C, γ, training dataset is used to train the SVM classifier, each chromosome is evaluated by fitness function

Step 6 : Termination criteria :When the termination criteria are satisfied, we evaluate a model using a full kdd99 data set; otherwise, return to step 3.

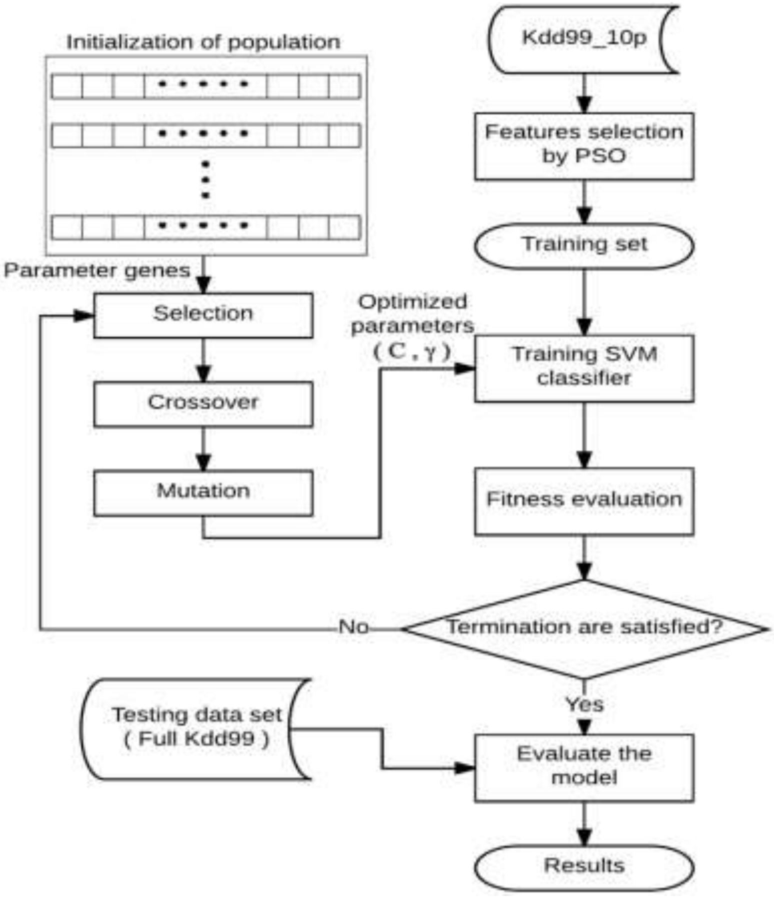
Initialization of population
Kdd99_10p
Features selection by PSO
Training set
Parameter genes
Selection
Optimized parameters (C,γ)
Training SVM classifier
Crossover
Mutation
Fitness evaluation
No
Termination are satisfied?
Yes
Testing data set (Full Kdd99)
Evaluate the model
Results

FIGURE 2: PROPOSED ALGORITHM

4. EXPERIMENT SETUP AND PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed model. All experiments were conducted on a calculation station 24 CPU Intel Core 2.13GHz, 48GB RAM, running under Linux CentOS 7. The implementation was coded using the Java language.

4.1. Data set

Cyber Systems and Technology Group of MIT Lincoln Laboratory [11] simulated LAN US Air Force LAN with multiple attacks and captured nine weeks TCPdump data. This database was first used for competitions kdd99, but since it has become the database test to the IDS's based on a behavioral approach. KDD Cup 1999 provided both the training dataset, it is called KDd99_10p.
Each connection record consists of approximately 100 bytes. This was converted into about 49 * 105 connection vectors each one contains 41 fields.
This database is collected by simulating attacks on different platforms such as Windows, Unix, etc ... Four gigabytes of raw data compressed TCP dump is transformed into five million connections files. The attacks are divided into four main categories: Denial of Service Attack (DOS), Probing, User to Root Attack (U2R), Remote to Local Attack (R2L).

DOI: 10.31695/IJASRE.2018.32724

4.1. Anomaly Detection Results

This section describes the obtained results from the experiment by applying the proposed algorithms on the data set kdd99. The performance of the proposed method of intrusion detection was evaluated on all KDD99 data set, 10% of the KDD99 data set were used for training the GA-SVM model after the features selection by PSO. Tables 2 illustrate the confusion matrix. this system achieves a top performance of up to 96,01% with a reasonable false alarm rate of 0,02% and a detection rate of 96,38%.

TABLE 2: CONFUSION MATRIX

Actual Class	Classified Class				
	Normal	DOS	Probe	R2L	U2R
Normal	713680	106761	26767	30811	15762
DOS	580	3878986	3784	6	14
Probe	519	6646	33797	3	137
R2L	4	0	17	923	182
U2R	0	0	0	21	31

Figure 3 shows the detection rate classified by attack, the proposed algorithm has detected 99,89%of DOS attacks whom are the most used by hackers. For Probe attacks a rate of 82,23% is correctly classified, 81,97% for R2L attacks and 59,62% for U2R, the low rate of detection U2R attacks can be explained by the insufficient number of data record-learning.

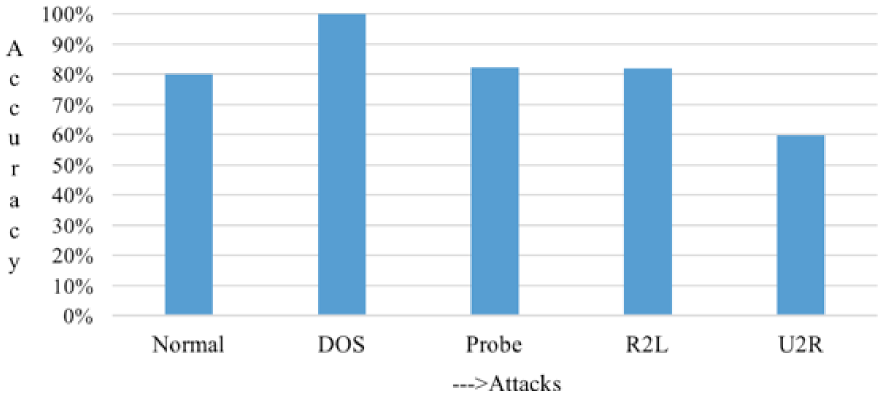

Figure 3: The accuracy rates per attack

To compare a detection rate, Figure 4 compares the proposed method with approaches that use only the entire KDD Cup 1999 dataset as a testing dataset because several researchers used only part of the KDD Cup 1999. The above results show that our approach enhances the performance of IDS. GA-SVM with PSO selection features is more reliable than state- of-the-art methods.

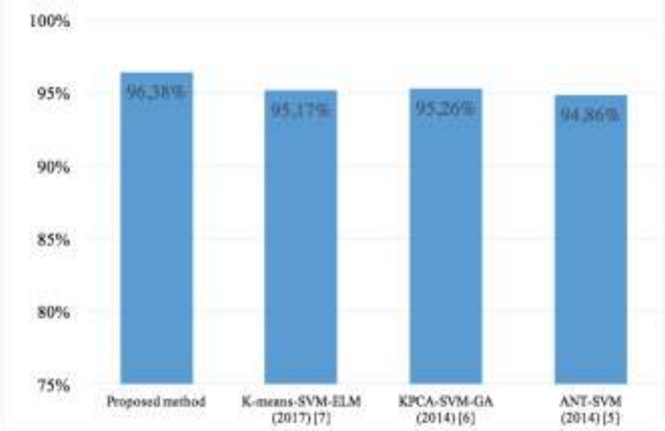

Figure 4: Comparison of proposed model with other methods by detection rate

The strengths of the proposed method are the highly improved detection accuracy compared with other methods because of the high reduction of original training dataset size which simplified the learning phase of the classifier and the SVM parameters optimization by GA.

5. CONCLUSION

In this paper, a Novel hybrid GA-SVM with PSO feature selection–based is proposed for intrusion detection. The proposed model is marked by a significantly better performance. RBF kernel function is used for improve the performance of SVM classification model, GA is used to select suitable parameters for SVM classifier and PSO is designated to select a features of the training dataset and provide new high-quality training datasets that can improve the overall performance of GA-SVM.
We also carried out comparisons of our method against other methods, and have shown a noticeable performance.
For future work, we want to develop more approaches to combine several machine learning techniques into one predictive model, using meta-algorithms, to increase the rate of detection of attacks.

ACKNOWLEDGEMENT

This work is supported in part by the High School of Technology, Moulay Ismail University Meknes, which has provided the calculation station where we have executed our experience.

REFERENCES

[1] J. P. Anderson, “Computer Security Threat Monitoring and Surveillance” Technical Report, James P. Anderson Company, Fort Washington, 1980.

[2] D.E. Denning, “An Intrusion-Detection Model”. IEEE Transactions on Software Engineering. February 1987, Vol. 13, pp. 222-232.

[3] S. Forrest, S. A. Hofmeyr, A. Somayaji and T. A. Longstaff, “A sense of self for Unix processes”, Proceedings of the 1996 IEEE Symposium on Security and Privacy, Oakland, California, USA. May1996, pp. 120-128.

[4] W. Lee, S.J. Stolfo and K.W. Mok, “A data mining framework for building intrusion detection models”, Proceedings of IEEE Symposium on Security and Privacy, Oakland, California, USA. May 1999, pp. 120–132.

[5] W. Feng, Q. Zhang, G. Hu, & J. X.Huang, “Mining network data for intrusion detection through combining SVMs with ant colony networks”, Future Generation Computer Systems, July 2014, vol. 37, pp 127–140.

[6] F. Kuang, W. Xu, & S. Zhang, A novel hybrid KPCA and SVM with GA model for intrusion detection. Applied Soft Computing Journal, May 2014, vol. 18, pp 178–184.

[7] L. A. Wathiq, A. O. Zulaiha, Z. A. Mohd , “Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-means for Intrusion Detection System”, Expert Systems with Applications, January 2017, vol. 67, Pages 296-303.

[8] J. Kennedy, R.C. Eberhart, “Particle swarm optimization”, Proceedings of of the 1997 IEEE International Conference on Neural Networks, November 1997, pp. 1942–1948.

[9] J. Kennedy, R.C. Eberhart. “A discrete binary version of the particle swarm algorithm”, Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation, IEEE International. October 1997 , pp. 4104–4109.

[10] C. Cortes, V. Vapnik, "Support vector networks", Machine Learning, September 1995, vol. 20,pp.273-297,.

[11] Holland, J. H. Adaptation in Natural and Artificial Systems. Cambridge, MA: MIT Press. Second edition (1992). (First edition, University of Michigan Press, 1975)

[12] M. Tavallaee, E. Bagheri, W/ Lu, and A. A. Ghorbani. “A Detailed Analysis of the KDD CUP 99 Data Set”, Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications, July 2009, pp. 1 – 6.

DOI: 10.31695/IJASRE.2018.32724