Application of AI in Network Intrusion Data Analysis and Intrusion Detection

[1]Raja Kumar Medapati , [2]Nadipalli Veda Vyas , [3]K. Jyothi , [4]Geeta Vijaya Kumar

Department of CSE,PRAGATI EngineeringCollege (Autonomous), Surampalem, A.P, India.

ABSTRACT

As new security intrusionsarise so does the demand for viable intrusion detection systems. These solutions must deal with huge data volumes, high speed network traffics and countervail new and various types of security threats. In this paper we combine existing technologies to construct an Anomaly based Intrusion Detection System. Our approach improves the Support Vector Machine classifier by exploiting the advantages of a new swarm intelligence algorithm inspired by the environment of microbats (Bat Algorithm). The main contribution of our paper is the novel feature selection model based on Binary Bat Algorithm with L´evy flights. To test our model we use the NSL-KDD data set and empirically prove that L´evy flights can upgrade the exploration of standard Binary Bat Algorithm. Furthermore, our approach succeeds to enhance the default SVM classifier and we obtain good performance measures in terms of accuracy (90.06%), attack detection rate (95.05%) and false alarm rate (4.4%) for unknown attacks.
INDEX TERMS: Intrusion Detection, SVM, Bat Algorithm, Binary Bat Algorithm, L´evy Flights.

INTRODUCTION

Many of our activities imply using the Internet (online payments, internet banking, social networks or searching for informations) and almost all government or private organizations store critical data over the networks. This increasing usage and growing speed of network connections has determined the proliferation of various threats. In this context security systems have become vital components. Despite the recent advances, security incidents are on the rise. IDS have become an indispensable component of almost every security infrastructure, mainly because they provide a wall of defense and resist external attacks effectively, where other traditional security systems cannot perform well. Intrusion Detection Systems (IDS) monitor the activities and events occurring in the systems and decide if these are intrusive actions or normal usage of the system. In general, IDS are classified on the basis of their data analysis technique as: misuse and anomaly detection. The misuse method is very accurate in detecting known attacks based on their signatures that are stored in the database. The anomaly detection approach automatically constructs a normal behavior of the systems. This latter method can

detect new attacks but, it can also generate many false alarms(Kukielka and Kotulski, 2014). Since 1980, when Anderson introduced the first IDS model, multiple techniques have been proposed to improve these systems. Many researchers have focused their attention on the anomaly based IDS and designed several approaches based on machine learning algorithms (decision trees, neural networks or genetic algorithms). There are many challenges to be regarded when implementing an IDS such as offering real-time responses with a high attack detection rate and a low false alarm rate. Also, the large number of features and difficulty to recognize the complex relationship between them makes classification a difficult task. In order to address these issues, we propose a Network Anomaly Intrusion Detection Model based on Support Vector Machines(SVM) with Swarm Intelligence(SI). Our approach has two main components: feature selection and enhanced SVM classifier. Support Vector Machine (SVM) has many advantages that make it a suitable solution for intrusiondetection, such as: good generalization performances and learning ability in high dimensional or noisy datasets. Furthermore, SVM does not suffer from local minima and its execution time runs fast. However, a draw-back of this classifier is that its performance depends on selection of the right parameters. Recently, swarm intelligence (SI) algorithms have attracted great interest, mainly because they are simple, flexible (can be applied to a variety of problems such as optimization, data mining and so forth) and robust (the algorithm will function even if some individuals fail to perform their tasks). Currently, there are a variety of SI algorithms such as: ant colony optimization, particle swarm optimization (PSO), artificial bee colony algorithm, firefly algorithm, cuckoo search or bat algorithm. These algorithms have been combined with SVM to contruct improved IDS models and in most cases are used for two optimization processes: feature selection and electing SVM parameters. Wang et. al. (Wang et al., 2009) present an IDS based on PSO-SVM. They used Binary PSO to determine the best feature subset and Standard PSO to seek for optimal SVM parameters. In 2010, a novel Artificial BeeColony(ABC)-SVM approach is proposed (Wang et al., 2010) and experiments on the KDDCup99 dataset proved that ABC- SVM can obtain better accuracy rate than PSO-SVM or GA-SVM. Pu et. al. (Pu et al., 2012) improve SVM parameters with the Ant Colony Algorithm by defining the input parameters as the ant's position. In this paper we use a relatively new metaheuristic to improve the SVM classifier, the Bat Algorithm (BA), which shows promising results (Yang, 2010a), (Yang and He, 2013). Furthermore, we introduce an innovative feature selection model that combines Binary Bat Algorithm (BBA) with Levy flights ´ and empirically demonstrate that it can outperform the standard BBA when coupled with SVM. The rest of this paper is organized as follows: first we introduce the algorithms that will be used to construct our IDS model. Next, we describe our approach, define the performance measures and show our test results. Also, we compare our model with other methods and indicate that it can obtain better results. Finally, the conclusions and future work

.

RELATED WORK

Support Vector Machine Support Vector Machine (SVM) is a binary supervised learning algorithm that searches for the optimum hyperplane to separate the two classes. This linear classifier conducts structural risk analysis

of statistical learning theory by selecting a number of parameters based on the requirement of the margin that separates the data points (Dua and Du, 2011b). The hyperplane is defined as f(x) = w· x + b, where w is the weight vector and b is the bias. Our classifier defines a hyperplane and linearly separates the two classes: anomaly and normal traffic. The points closest to the hyperplane are called support vectors and the distance between them is called the margin. On the other hand, the dataset is not always linearly separable. This issue is solved by introducing a slack variable and defining a soft margin. Furthermore, using kernel functions, we can transform the nonlinear SVM into a linear problem by mapping the dataset into a higher-dimensional feature space. For our model we will use SVM with radial basis function (RBF), where the kernel function is: K(xi ,x) = exp(− 1 2σ2 kxi −xk 2) (1) The RBF kernel function is a good solution because it has fewer controllable parameters and an excellent nonlinear forecasting performance. In the following we define the SVM controllable parameters and explain their influence. • The Regularization Parameter (C) - controls the ”flexibility” of the hyperplane’s margins. In other words, smaller C allows softer-margins thus, permits greater errors. Larger C produces a more accurate model with harder margins but, the generalization performance of the classifier is worse. • Kernel Parameter (σ) - is the constant variable from the kernel function. This parameter reflects the correlation among support vectors that define the hyperplane and may cause overfitting or underfitting of the classifier.

Bat Algorithm Bat Algorithm (BA) was developed by Yang in 2010 (Yang, 2010a) and it was inspired by the echolocation of bats. These microbats emit a loud sound pulse and change their pulse rate as the obstacle or prey is closer. In order to define a smart bat algorithm, three generalization rules have been formulated:

• All bats use echolocation to approximate the difference between an obstacle and a prey and sense distance. • Bats fly randomly and their movement is defined by their position in space (xi) and velocity (vi). These parameters are computed based on a varying wavelength (λ), frequency (f reqmin) and loudness (A0) to search for prey. Moreover, bats can modify the frequency of their emitted pulses and the rate of pulse emission (r ∈ [0,1]), depending on the closeness of their target. • The loudness can vary in multiple ways, but we assume that it varies from a large value (A0) to a minimum constant value (Amin). BA is a swarm intelligence algorithm which performs searches using a population of agents. For SVM parameter selection, BA will search for the best C and σ based on the accuracy of SVM. Each agent i has a current position xi = (xi,1,xi,2,...,xi,d) t and a EnhancedIntrusionDetectionSystemBased onBatAlgorithm-supportVectorMachine 185

$$freq_i = freq_{min} + (freq_{max} - freq_{min}) \cdot \beta \quad (2)$$

$$v_{i,j}^{t} = v_{i,j}^{t-1} + (x_{i,j}^{t-1} - x\_best_j) \cdot freq_i \quad (3)$$

$$x_{i,j}^{t} = x_{i,j}^{t-1} + v_{i,j}^{t} \quad (4)$$

where β ∈ [0,1] is a random vector drawn from a uniform distribution. As stated earlier, the bat will decreases his loudness (Ai) and increase his pulse emission rate (ri) when he is closer to the target:

$$A_i^{t+1} = \alpha \cdot A_i^{t} \quad (5)$$

$$r_i^{t+1} = r_i^{0} \cdot [1 - e^{-\gamma \cdot t}] \quad (6)$$

where α (0 < α < 1) and γ (γ > 0) are constants. At each iteration, the fitness value is improved, while A → 0 and r → r0. In order to ameliorate the variability of the discovered solutions, Yang uses random walks to generate new solutions:

$$x_{new} = x_{old+\delta} \cdot A^{*}{}_{t} \quad (7)$$

where δ ∈ [−1,1] is a random number and A ∗ t is the average loudness of all bats at iteration t. In some ways we can state that BA is similar to the popular PSO. The solution is denoted by the position of the particle, each individual from the swarm has its own position and velocity which are updated according to their fitness value. There are also some significant differences, as BA uses random walks for exploration (or diversification) and varies the loudness and pulse rate in order to exploit the solution. For PSO, exploitation is controlled by the use of the global best and individual best solutions, while exploration is done using two learning parameters(Yang, 2012). 2.3 Feature Selection Feature selection methods can be divided into : scalar methods (are less complex and select features individually) and vector methods (select a subset of features based on a mutual relation between features)(Dua and Du, 2011a). In the following we introduce the Binary Bat Algorithm and explain how it can be adapted to construct a vector approach feature selection. 2.3.1 Binary Bat Algorithm The Binary Bat Algorithm(BBA) (Mirjalili et al., 2013) is a modified version of BA that describes the bat's motion in a d-dimensional binary space. Therefore, the position of a bat is defined as a vector of binary coordinates and the bat can move across the corners of a hypercube. Given bat i, its coordinates are computed using a sigmoid function as follows: xi, j = (−1 if S(vi, j) > δ 0 otherwise (8) Where, the sigmoid function is : S(vi, j) = 1 1+e −vi, j (9) and δ ∈ U(0,1). Hence, the bat's position can be seen as a string of binary numbers. For the feature selection process we want to determine the best feature subset that will enhance the performances of our classifier. In order to adapt BBA for feature selection we can consider the bat's position as the subset of features and the bat's coordinates as the presence (if the coordinate is one) or absence(if the coordinate is zero) of a feature. The fitness function for BBA will be the accuracy of SVM after it has been trained with the subset of features represented by the bat's position. In order to elevate the exploration of BBA we use Levy flights ´ for randomization. Levy flights ´ are a random walk whose step length is drawn from a L´evy distribution. Recent studies have shown that L´evy flights can better searches in uncertain environments. L´evy flights have many applications and have been observed among foraging pattern of spider monkeys or albatrosses (Yang, 2010b). This distribution has been combined with other computational intelligence algorithms, such as : cuckoo search(Yang and Deb, 2009), firefly (Yang, 2009) or bat algorithm (Xie et al., 2013). Here we exploit this distribution with BBA for feature selection. Therefore, we substitute equation (7) with:

$$x_{new} = x_{old} + t - \eta \cdot A^{*}{}_{t} \quad (10)$$

where t −η is the L´evy flights distribution and 1 < η ≤ 3 is a constant

PROPOSED MODEL

Our model has three main stages : first we apply BBA with L´evy flights (BBAL) to determine the best subset for SVM, next we use BA to determine the best parameters for SVM. Finally, SVM detects network attacks using the best parameters calculated above. 3.1 Data Set In this paper we experimented with the NSL-KDD data set, which contains network attacks. We chose this dataset because it is publicly available and is an improved

version of KDD-Cup (Tavallaee et al., 2009). The simulated attacks from NSL-KDD fall into one of the following categories: Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L) or Probing. Furthermore, there are two different groups of files: training and testing. It is important to mention that test data includes attack types not in the training data and therefore it will allow us to evaluate the classifier for unknown attacks. We randomly select 9,500 feature samples from the 20% training file and 4,500 records from the test file. Each record from the dataset has 41 features and is labeled as either normal or an attack. These features can be classified into three categories: content based (13 features), connection based (9 features) and time based (19 features). In order to improve the prediction ability of the classifier, we convert the symbolic values into integer values, as follows: protocol type ∈ [0,2] , service ∈ [0,69] and f lag ∈ [0,10]. We also replace the class label with 0 for normal and 1 for attack. 3.2 Evaluation To evaluate our model we find the best subset of features with BBAL and compare it with BBA. Also, we train the SVM classifier with BA using the determined feature subset and compare it with PSO. This comparison is relevant because PSO is a popular swarm intelligence algorithm that has been widely used for optimization problems. 3.2.1 Performance Measures We estimate the effectiveness of our IDS model by calculating three performance measures: attack detection rate (ADR) (shows the model's capability of detecting attacks), false alarm rate (FAR) (measures how many false alarms the model generates) and accuracy (reveals the model's capability of raising proper alarms). 3.2.2 Model Setup All experiments were performed using an Intel Core 2 Duo 2.8 GHz processor with 2 GB of RAM under Ubuntu 10.04.4. We implemented the Swarm Intelligence algorithms (BBA, BBAL, BA and PSO) in Java on eclipse. For the SVM classifier we used Weka version 3.6.10 (Hall et al., 2009). BBAL and BBA

3.2.3 Results and Analysis Tables 1 and 2 indicate the best subset of features determined with BBAL-SVM and BBA-SVM, having parameter C = 1.0 and σ = 0.5. BBAL obtains the smallest number of features, reducing them by 43.9%. BBA is more vulnerable to local optima than BBAL but, it succeeds to simplify the number of features. To evaluate these subset we use the training dataset to build the model and the test dataset to evaluate it. Also, we perform a 3-fold cross validation for the training dataset, in order to show the model's accuracy for known attacks. Results from Table 3 reveal the BBA-SVM subset gives slightly better performances when compared to the initial dataset, leading SVM to a higher classification rate, attack detection rate and lower number of false alarms. On the other hand, the BBAL subset, having a smaller number of features, succeeds to enhance the classification abilities of SVM and offers betterresults than BBA. Furthermore, we also try to improve the SVM classifier by selecting the proper input parameters. For this we use two swarm intelligence algorithms (PSO and BA) that will search for the best solution. In other words, each individual of the swarm will attempt to acquire a position, represented by the two parameters, that will bring them a higher accuracy for the classifier. The comparison between PSO and BA for the two subsets is shown in Tables 4 and 5. We can observe that both PSO and BA manage to boost up the performance of the classifier by electing the adequate parameters for SVM. As expected, the enSECRYPT2014-InternationalConferenceonSecurityandCryptography 188

CONCLUSIONS

In this paper we proposed a new NIDS model that combines SVM with a recent swarm intelligence algorithm, the Bat Algorithm. The main contribution of this paper is the novel feature selection method (BBAL) that succeeds to reduce the number of attributes from the dataset while improving the predictive accuracy, detection rate and false alarm rate of the SVM classifier. To evaluate the effectiveness of the proposed model we use the NSL-KDD network intrusion benchmark and compare it with the popular PSO for our two subset of features. We showed that BBAL can upgrade BBA for feature selection but, only when combined with SVM. Therefore, our future work will focus on combining BBAL with other classifiers and comparing it to other feature selection approaches in order to range its quality.

ACKNOWLEDGEMENTS

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007- 2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/132395.

REFERENCES

Dua, S. and Du, X. (2011a). Classical machine-learning paradigmsfor data mining. In Data Mining and Machine Learning in Cybersecurity, pages 23–56. Auerbach Publications Taylor and Francis Group. Dua, S. and Du, X. (2011b). Machine learning for anomaly detection. In Data Mining and Machine Learning in Cybersecurity, pages 85–114. Auerbach Publications Taylor and Francis Group. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. SIGKDD Explor. Newsl., 11:10– 18. Kukielka, P. and Kotulski, Z. (2014). New unknown attack detection with the neural network-based ids. In The State of the Art in Intrusion Prevention and Detection, pages 259–284. Auerbach Publications. Mirjalili, S., Mirjalili, S., and Yang, X.-S. (2013). Binary bat algorithm. Neural Computing and Applications, pages 1–19. Pu, J., Xiao, L., Li, Y., and Dong, X. (2012). A detection method of network intrusion based on svm and ant colony algorithm. In Proceedings of the National Conference on Information Technology and Computer Science, pages 153–156. Atlantis Press. Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In Proceedings of the IEEE Symposium on Computational Intelligence in Security and Defense Applications, pages 1–6. IEEE. Wang, J., Hong, X., and R. Ren, T. L. (2009). A realtime intrusion detection system based on pso-svm. In Proceedings of the International Workshop on Information Security and Application, pages 319–321. ACADEMY PUBLISHER. Wang, J., Li, T., and Ren, R. (2010). A real time IDSs based on artificial bee colony-support vector machine algorithm. In Proceedings in the International Workshop on Advanced Computational Intelligence, pages 91–96. IEEE. Xie, J., Zhou, Y., and Chen, H. (2013). A novel bat algorithm based on differential operator and l´evy flights trajectory. Computational Intelligence and Neuroscience, 2013. Yang, X.-S. (2009). Firefly algorithm, l´evy flights and global optimization. In Proceedings of the SGAI International Conference on Artificial Intelligence, pages 209–218. Yang, X.-S. (2010a). A new metaheuristic bat-inspired algorithm. In Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), volume 284 of Studies in Computational Intelligence, pages 65–74. Springer Berlin Heidelberg. Yang, X.-S. (2010b). Random

walks and l´evy flights. In Nature-Inspired Metaheuristic Algorithms Second Edition, pages 11–20. Luniver Press. Yang, X.-S. (2012). Swarm-based metaheuristic algorithms and no-free-lunch theorems. In Theory and New Applications of Swarm Intelligence. InTech. Yang, X.-S. and Deb, S. (2009). Cuckoo search via l´evy flights. In Proceedings of the World Congress on Nature & Biologically Inspired Computing, pages 210– 214. IEEE