PROGRESS OF MACHINE LEARNING IN THE FIELD OF INTRUSION DETECTION SYSTEMS

Ouafae Elaeraj and Cherkaoui Leghris

L@M, RTM Team, Faculty of Sciences and Techniques Mohammedia,
Hassan II University of Casablanca, Morocco

ABSTRACT

With the growth in the use of the Internet and local area networks, malicious attacks and intrusions into computer systems are increasing. Implementing intrusion detection systems have become extremely important to help maintain good network security. Support vector machines (SVMs), a classic pattern recognition tool, have been widely used in intrusion detection. They can handle very large data with high efficiency, are easy to use, and exhibit good prediction behavior. This paper presents a new SVM model enriched with a Gaussian kernel function based on the features of the training data for intrusion detection. The new model is tested with the CICIDS2017 dataset. The test proves better results in terms of detection efficiency and false alarm rate, which can give better coverage and make detection more efficient.

KEYWORDS

Intrusion detection System, Support vector machines, Machine Learning.

1. INTRODUCTION

The majority of intrusion detection systems are software defense systems. The main functions of an IDS are to monitor events occurring in a computer system or network, analyze system events, detect activities, and raise an alarm if an intrusion is detected. An IDS can be divided into three functional components [3]: an information source, an analysis engine and a decision maker. Figure 1 shows the relationship between these three components, the Internet, and the protected systems.

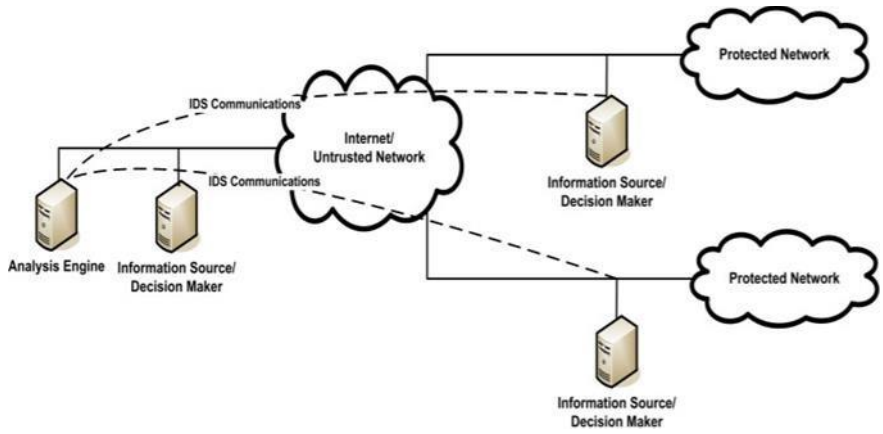


Figure 1. Relation between IDS components.

In the first IDS component, the information source has the task of collecting and pre-filtering all relevant and necessary data to unmask intruders. During the monitoring period, the information source provides a stream of event records for analysis. This component works as an event generator. It detects and monitors different data sources and generates well-formatted event data suitable for further analysis by an analysis engine.

IDS, using anomaly detection, are able to detect unknown attacks, but with the risk of progressive distortion of the profile by repeated attacks. In this context, the creation of anomaly- accurate IDS is of major interest to be able to identify still unknown attacks. Machine learning models can be a solution to create IDS that can be deployed in real computer networks. First of all, an optimization method composed of three steps is proposed to improve the quality of the detection: 1/ data augmentation to rebalance the datasets, 2/ parameter optimization to improve model performance, and 3/ assemble learning to combine the results of the best models. This paper proposes a new SVM model to have a better coverage and make the detection more efficient. The rest of the paper is organized in the following sections: In Section 2, we briefly describe the various existing research of using SVM in IDS. In section 3, we study the proposed solution. The results are described in section 4. At the end, we summarize the paper and address some prospects.

1.1. Machine Learning

Machine Learning is a form of artificial intelligence that allows computers to learn without being explicitly programmed to do so. It is a scientific discipline that deals with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as sensor data or databases. And also, a data analysis method to automate the development of analytical models.

Machine Learning algorithms are not new, but it is only recently that it has become possible to apply complex mathematical calculations to Big Data at an ever-increasing speed. Machine Learning is now used in many fields, such as the development of autonomous vehicles, online recommendation systems like those of Netflix and Amazon, analysis of customer sentiment, or fraud detection.

The resurgence of interest in Machine Learning is related to the same factors that have been driving attention around data mining and analytics technologies. Data is increasingly plentiful and diverse, computing power is cheaper than ever, and data storage is now affordable.

ML is used in many applications, including search engines, medical diagnostics, text and handwriting recognition, load forecasting, marketing and business diagnostics, and more. In 1994, ML was first used for Internet stream classification in the context of intrusion detection [1]. This was the beginning of a large body of work using ML techniques in Internet traffic classification. In this paper, the SVM algorithm was chosen because it is the best, in terms of accuracy and processing time, after KNN and decision tree [2].

1.2. Support Vector Machines

Generally, Support Vector Machines (SVM) are a class of learning algorithms initially defined for discrimination, i.e. the prediction of a binary qualitative variable. They were then generalized to the prediction of a quantitative variable. In the case of the discrimination of a dichotomous variable, they are based on the search for the are based on the search for the optimal margin hyperplane which, when possible, correctly possible, correctly classifies or separates the data while being as far as possible from all the observations. The principle is therefore to find a

classifier, or a discrimination function, whose generalization capacity (predictive quality) is as high as possible.

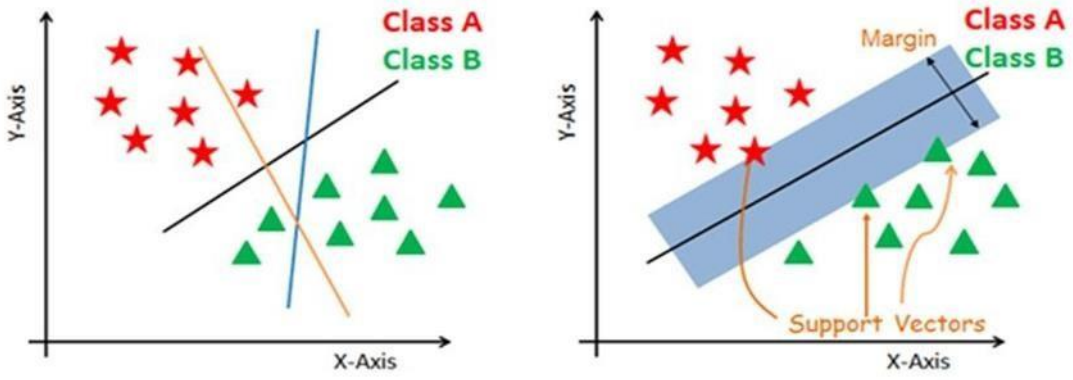
This approach follows directly from Vapnik's work in learning theory since 1995. It focused on the generalization (or prediction) properties of a model by controlling its complexity. The founding principle of SVMs is precisely to integrate the control of complexity into the estimation, i.e. the number of parameters which is associated in this case with the number of support vectors. Recently, it has also been applied to information security for intrusion detection. Support Vector Machine has become one of the anomaly intrusion detection techniques because of their good generalization nature.

1.2.1. Operation of the SVM

The main objective is to separate the data set in the best possible way. The distance between the two closest points is called the margin. The purpose is to select a hyperplane with the maximum possible margin between the support vectors in the given dataset (Figure 2). SVM searches for the maximum marginal hyperplane in the following steps:

1. Generate hyperplanes that separate the classes. Figure 2 (a) shows three black, blue and orange hyperplanes. Here blue and orange have a higher classification error, but black correctly separates the two classes.

2. Select the right hyperplane with the maximum separation of the closest data points, as shown in the figure 2(b).



(a) **(b)**

Figure 2. Operation of the SVM

1.3. IDS using SVM

There are several reasons why we use SVMs for intrusion detection. The first is speed: in real time, performance is of primary importance for intrusion detection systems. Any classifier that can potentially perform "fast" is worth considering.

The second reason is scalability: SVMs are relatively insensitive to the number of data points and the complexity of classification does not depend on the dimensionality of the

feature space, so they can potentially learn a larger set of models and thus be able to scale better than neural networks.

2. RELATED WORK

Intrusion detection has a long history, dating back to the work of Anderson (1980) [3]. Since long, various discrimination techniques have been proposed, ranging from support vector machines. Several complete systems have been built and operated on real computer systems. However, despite more than 25 years of research, the topic remains popular, partly because of the rapid development of information processing systems and the consequent discovery of new vulnerabilities, but also because of the fundamental difficulties in obtaining an accurate report of an intrusion.

In paper [4], the authors propose a genetic algorithm (GA) to improve the support vector machine (SVM) based intrusion detection system (IDS). As known, SVM is a relatively new classification technique and has shown superior performance to traditional learning methods in

many applications. The fusion of GA and SVM has been used in this article to fortify the overall performance of SVM-based IDS. Through this fusion, SVM-based IDSs not only select the "optimal parameters" for SVMs but also an "optimal set" from the feature set.

An anomaly-based IDS using a genetic algorithm and a support vector machine (SVM) with a novel feature selection method is also proposed in [5]. The model adopts a feature selection method based on the genetic algorithm with a change in the fitness function that sums up the size of the data, increases the detection of true positives and simultaneously decreases the detection of false positives. In addition, the computational time for learning will also be reduced remarkably. The results show that the proposed method can simultaneously achieve high accuracy and low false positive rate (FPR). This study proposes a method that achieves more stable features compared to other techniques. The proposed model is experimented and tested on the KDD CUP 99 and UNSW-NB15 datasets.

Network traffic in cloud computing is characterized by a large volume of data, with high levels of redundancy. The efficient correlation-based feature selection (ECOFS) approach, proposed in [6], can handle linear and non-linear dependent data and eliminates redundant and irrelevant features. Its effectiveness has been examined using an intrusion detection system. A Libsvm- IDS has been designed to operate using the features selected by the proposed ECOFS algorithm. The results of the evaluation show that the ECOFS algorithm selects the smallest number of features, resulting in the lowest computational cost for the Libsvm-IDS, with better performance. In fact, the algorithm has achieved greater precision.

In order to optimize the training procedure of SVM-based intrusion detection systems and reduce the time consumption, a GPU-based SVM intrusion detection method is proposed in [7]. During the simulation experiments with the KDD 1999 Cup data, the GPU-based parallel computing model is adopted. The results of the simulation experiments show that the time consumption in the IDS training procedure is reduced, and the performance of the IDS is maintained as usual.

Although IDSs have been in development for many years, the large number of return alerts make system maintenance inefficient for managers. In this article [8], the use of RST (Rough Set Theory) and SVM has been blown to detect intrusions. First, RST is used to pre-process data and reduce dimensions. Then, the features selected by RST are sent to the SVM model to learn and test respectively. This method is effective in reducing the spatial density of the data. The

experiments compare the results with different methods and show that RST and SVM scheme can improve the false positive rate and accuracy.

The paper [9] proposes a Factor Analysis based Support Vector Machine (FA-SVM) algorithm to develop efficient IDSs using the popular statistical technique called factor analysis (FA). To design more effective and efficient IDSs, it was essential to select the best classifiers. This work is performed on the Knowledge discovery dataset and data mining to perform tests. The performance of this approach was compared to existing approaches such as principal component analysis (PCA) using SVMs, as well as classification with SVMs itself without feature selection. The results prove that the proposed method improves the detection of intrusions and intrusions, in terms of calculating false positive rates.

Wireless fidelity (WiFi) is a widely used test area due to its mobility in the presence of the main disadvantage of securing the network. Several attempts to secure 802.11 result, in inadequate security mechanisms, that this technology is vulnerable to various attacks and intrusions. The paper [10] proposes a Normalized Gain for MAC Intrusion (NMI)-based IDS to significantly improve the performance of the IDS. The proposed NMI consists of two primary components OFSNP and DCMI. The first component is the optimal feature selection using NGand PSO (OFSNP) and the second one is the detection and categorization of 802.11 MAC intrusions (DCMI) using the SVM classifier. Proposed NMI achieves a better trade-off between detection accuracy and learning time. The experimental results show that NMI accurately detects and classifies 802.11 specific intrusions and also reduces false positives.

3. PROPOSED SOLUTION

The CICIDS2017 dataset contains the most recent common benign attacks that resemble real world data (PCAP). It also includes the results of network traffic analysis using CICFlowMeter with flows labeled by timestamp, source and destination IP addresses, source and destination ports, protocols, and attacks (CSV files). For this dataset, we constructed the abstract behavior of 25 users based on HTTP, HTTPS, FTP, SSH and email protocols.

The proposed algorithm uses a set from CICIDS2017 to select important features. The reduced feature CICIDS2017 dataset is then used for training and designing a detection model on the SVM classifier.

We divide the data into two classes: normal and attack, where the implemented attacks include Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet and DDoS. The goal of our SVM implementation is to separate the normal and attack models. In our case, we will take 3 days traffic to get accurate results by keeping the first day as training set and the rest of the csv files as test set:

- **Monday July, 3rd 2017**

Benign (Normal human activity)

- **Tuesday July, 4th 2017**

Force brute
FTP-Patator (9h20 - 10h20) SSH-Patator (14h00 - 15h00)

- **Wednesday July, 5th 2017**

DoS / DDoS

DoS slowloris (9h47 - 10h10) DoS Slowhttptest (10h14 - 10h35) DoS Hulk (10h43 - 11h) DoS GoldenEye (11h10 - 11h23).

The procedure of the solution note is as follows:

1- Importing libraries ;
2- Data import ;
3- Data selection and indexing ;
4- Data pre-processing :

- Data pre-processing involves dividing the data into training and test sets ;

5- Training the algorithm :

- Unsupervised outlier detection with a kernel (Specifies the type of kernel to use in the algorithm. It can be "linear", "poly", "rbf", "sigmoid", "precalculed". If none is specified, "rbf" will be used.) of parameters gamma(kernel coefficient for rbf and poly, if gamma is 0.0 then 1 / n_features will be taken) ;
- Detection of the soft limit of the sample set.

6- Evaluation of the algorithm :

- Confusion matrix, precision, recall are the most commonly used measures for classification tasks ;

4. TESTING

In our first experiment, we will take as training data source the Monday traffic, which is a benign traffic, and a Tuesday test set, which is composed of attacks and normal activities, and we will apply our solution with the parameter's gamma = 0.0005, kernel = 'rbf', nu=0.001. The results show that the precision reaches up to 97 % (Figure 3).

	precision	recall	f1-score	support
0	0.97	1.00	0.98	431813
1	0.12	0.01	0.01	13832
avg / total	0.94	0.97	0.95	445645

Figure 3. Results of the first experiment with parameters (gamma = 0.0005, kernel = 'rbf', nu=0.001)

In the second experiment, we will take as training data source the Monday traffic which is a benign traffic and a test set of Wednesday which consists of attacks plus normal activity, and we will apply our solution with the parameter's gamma = 0.0005, kernel = 'rbf', nu=0.001. The results show that the precision reaches up to 99 % (Figure 4).

	precision	recall	f1-score	support
0	0.73	1.00	0.84	439683
1	0.99	0.34	0.51	251723
avg / total	0.82	0.76	0.72	691406

Figure 4. Results of the first experiment with parameters (gamma = 0.0005, kernel = 'rbf', nu=0.001)

5. RESULTS

Table 1 shows the results of our first experiment with different values of nu and gamma.

Table 1. The results of the first experiment with different values of nu and gamma.

Scenario	nu	Gamma	avg / total	Precision	Recall	f1-score	TP	FP	TN	FN
Monday/Tuesday	0,01	0,05	0	0,98	0,99	0,99	6877	5010	426803	6955
			1	0,58	0,5	0,53				
				0,97	0,97	0,97				
	0,05	0,1	0	0,99998	0,93782	0,9679				
			1	0,33983	0,99928	0,50718	13822	26851	404962	10
				0,97949	0,93973	0,9536				
	0,1	0,05	0	0,99992	0,89304	0,94346				
			1	0,23005	0,99769	0,37389	13800	46186	385627	32
				0,97602	0,89629	0,37389				
	0,03	0,2	0	1	0,96391	0,98163				
			1	0,47025	1	0,63969	13832	15582	416231	0
				0,98356	0,96503	0,97101				
	0,3	0,001	0	1	0,70609	0,82773				
			1	0,09828	1	0,17897	13832	126912	304901	0
				0,97201	0,71522	0,8076				
	0,07	0,02	0	0,99347	0,92186	0,95632				
			1	0,24944	0,81073	0,38151	11214	33742	398071	2618
				0,97037	0,91841	0,93848				

Table 2 shows the results of the second experiment with different values of nu and gamma.

Table 2. The results of the second experiment with different values of nu and gamma.

Scenario	nu	Gamma	avg / total	Precision	Recall	f1-score	TP	FP	TN	FN
Monday /Wednesday	0,01	0,05	0	0,85385	0,98627	0,9153				
			1	0,9671	0,70514	0,8156	177499	6038	433645	74224
				0,89508	0,88391	0,879				
	0,03	0,2	0	0,99204	0,95918	0,97533				
			1	0,9326	0,98655	0,95882	248338	17949	421734	3385
				0,9704	0,96914	0,96932				
	0,1	0,02	0	0,87375	0,905	0,8891				
			1	0,82301	0,77159	0,79647	194227	41768	397915	57496
				0,85528	0,85643	0,85538				
	0,05	0,1	0	0,91202	0,94448	0,92796				
			1	0,8966	0,84085	0,86783	211661	24411	415272	40062
				0,9064	0,90675	0,90607				
	0,05	0,05	0	0,87976	0,94588	0,91162				
			1	0,88392	0,88337	0,88139	194880	23794	415889	56843
				0,88392	0,88337	0,88139				
	0,03	0,05	0	0,87156	0,96275	0,91489				
			1	0,92038	0,75219	0,82783	189343	16380	423303	62380
				0,88933	0,8609	0,88319				

The results are significant, we get up to 98% accuracy with a zero false negative value, we do not forget that the data is voluminous Monday (11G), Tuesday (11G) and Wednesday (13G) including attacks : Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet and DDoS. In addition, we notice that when we increase the value of Gamma, we have more efficient precision and a lower false negative rate.

6. CONCLUSION

In order to improve the efficiency of the carrier vector machine (SVM) based intrusion detection system (IDS), we have conducted a large number of experiments to measure the performance of carrier vector machines in the intrusion detection, using CICIDS2017 data for intrusion assessment. SVMs provide very accurate performance (99% and above).

Our future work is to apply new SVM optimization techniques in SNORT fusion to improve attack detection in computer network security.

REFERENCES

[1] J. Frank, “Machine learning and intrusion detection: Current and future directions”, in Proceedings of the National 17th Computer Security Conference, Washington, October 1994 ;

[2] Elaeraj, Ouafae & Cherkaoui, Leghris & Renault, Éric, “Performance Evaluation of Some Machine Learning Algorithms for Security Intrusion Detection”, in the Machine Learning for Networking Third International Conference, Paris, 2020 ;

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

[4] Kim, Dong Seong, Nguyen, Ha-Nam & Park, Jong, “Genetic Algorithm to Improve SVM Based Network Intrusion Detection System”, in the 19th International Conference on Advanced Information Networking and Applications, Taiwan, 2005 ;

[5] H. Gharaee and H. Hosseinvand, “A new feature selection IDS based on genetic algorithm and SVM”, in the 8th International Symposium on Telecommunications (IST), United States, 2016 ;

[6] W. Wang, X. Du and N. Wang, “Building a Cloud IDS Using an Efficient Feature Selection Method and SVM”, in IEEE Access, vol. 7, pp. 1345-1354, 2019 ;

[7] Xia, Yong & Shi, Zhi & Zhang, Yu & Dai, Jian, “A SVM intrusion detection method based on GPU”, in the international Conference on Applied Mechanics, Mechatronics and Intelligent System, China, 2014 ;

[8] Chen, Rung-Ching, Cheng, Kai-Fan & Hsieh, Chia-Fen, “Using Rough Set and Support Vector Machine for Network Intrusion Detection”, in the International Journal of Network Security & Its Applications, 2010 ;

[9] P Indira Priyadarsini, I Ramesh Babu, “Building Efficient Intrusion Detection System Using Factor Analysis and Support Vector Machines”, in the international journal of engineering research & technology (IJERT) Volume 03, Issue 04, April 2014 ;

[10] Murugan, Kavitha & M, Usha, “Anomaly Based Intrusion Detection for 802.11 Networks based on Optimal Feature Selection using SVM Classifier”, Springer DOI: 10.1007/s11276016-1300-5, Wireless Networks,2016.

AUTHORS

Ouafae El AERAJ, aged 26, Research student in network security at the Faculty of Sciences and Techniques Mohammedia, Hassan II University of Casablanca, Morocco.

Cherkaoui LEGHRIS has a PhD in computer sciences from the Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Mohammed V University of Rabat, Morocco, in 2007. On 2003, He had the diploma of Higher Studies in ENSIAS and before the degree in Applied Computer Science from the Cadi Ayyad University of Marrakech. He is currently working as Professor of Higher Education at the Faculty of Science and Techniques, Hassan II University of Casablanca. He is responsible of various modules in communication networks domain.

He conducts research on networking at L@M Laboratory, RTM team. His main research focus on the IoT networks based across IPv6 protocol, multi-access network technologies and IT security. He is also an active member in the Moroccan Internet SOCiety organization, which works for openness and accessibility of the Internet for any.