Learning a Secure Classifier against Evasion Attack

Zeinab Khorshidpour, Sattar Hashemi, Ali Hamzeh
Shiraz University, Department of Electronic and Computer Engineering
Shiraz, Iran
Email: zkhorshidpur@cse.shirazu.ac.ir

***Abstract*—In security sensitive applications, there is a crafty adversary component which intends to mislead the detection system. The presence of an adversary component conflicts with the stationary data assumption that is a common assumption in most machine learning methods. Since machine learning methods are not inherently adversary-aware, it necessitates to investigate security evaluation of machine learning based detection systems in the adversarial environment. Research in adversarial environment mostly focused on modeling adversarial attacks and evaluating impact of them on learning algorithms, only few studies have devised learning algorithms with improved security. In this paper we propose a secure learning model against evasion attacks on the application of PDF malware detection. The experimental results acknowledge that the proposed method significantly improves the robustness of the learning system against manipulating data and evasion attempts at test time.**

***Keywords*-Machine learning; Adversarial classification; Adversary; Evasion attack; Robustness.**

I. Introduction

Machine learning techniques have been adopted in security sensitive applications like malware detection, spam filtering, botnet detection, and network intrusion detection to confront the problem of zero day attacks, since traditional security systems were not able to detect before unseen attacks, i.e. zero day attacks [1], [2]. Detection in these applications are known as adversarial classification. In this context, there is an adaptive and crafty adversary component which intends to mislead the detection system. The existence of this adversary component is due to inherent adversarial nature of security sensitive applications. The use of machine learning techniques in the adversarial environment introduces new challenges; because presence of the adversary conflicts with the stationary data assumption, that is assumed in the most machine learning methods.

Moreover, this assumption is a basis for performance evaluation techniques like cross-validation, bootstrapping, and empirical risk minimization principle. Evasion attack is a typical scenario caused by adversary in the adversarial environment. This attack aims to evade a learning classifier by increasing its false negative rates at test time. Moreover, it modifies the distribution of malicious samples of test data to resemble legitimate samples. For example, in PDF malware detection, malicious PDF files under this attack mimic the structure of legitimate PDF samples. Spammer as an adversary component tries to deceive classifier through bad word insertion like misspelling spammy word ”cost” as ”c0st”.

Considering the aforementioned nature of security applications, a new line of research has emerged in order to investigate the challenge of exploiting machine learning in an adversarial environment [3]. Machine learning techniques are not well established to address the adversarial nature of security applications. Thus, it is important to evaluate robustness of machine learning based detection system in the presence of adversaries.

In this paper, we propose a robust learning model against evasion attack on the application of PDF malware detection. There are different reasons that potentially can increase infection of PDF file format. The most important reason is due to the flexibility of logical structure of PDF format. The attackers can embed their malicious JavaScript, ActionScript, binary codes and other type of attacks into the PDF file. In addition, the attackers employ self-protection techniques like encryption and obfuscation to hide their malicious functionality. Hence, attackers are able to simply devise polymorphic attacks. This type of attack constitutes a family of malware with the same functionality but with different appearances which are hard to detect and fast to deploy. Researches used the machine learning methods to confront the attack polymorphism problem. The previous studies have shown malicious PDFs can be detected by their extracted structural information [4], [5], [6], [7]. However, the robustness of such detection systems are questioned where the malicious PDF files mimic the structure of legitimate PDF samples [4]. The main contributions of this paper are as following.

- First, we propose a robust learning model to tackle performance degradation due to adversary attack.
- Second, we address an adversarial model named as evasion attack on the assessment of SVM classifier and the proposed learning model at test time to investigate security evaluation of them.
- Third, we show experimentally that the proposed learning model effectively preserves its robustness under evasion attack.

The reminder of paper is organized as follows. Section 2 reviews previous work. Section 3 highlights the background

2375-9259/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDMW.2016.46

of this research. Section 4 presents a robust classifier against the evasion attack. Section 5 addresses the security evaluation of learning system. In Section 6, evaluation results are discussed. Finally, Section 7 describes next steps and concludes the paper.

II. Related Work

Data stationarity is the common assumption in the most machine learning techniques. According to this assumption training data in which the model is developed and test examples on which model is used have the same distribution. However, in the real world applications of machine learning train and test data may follow different distributions. This is a fundamental problem in the security sensitive applications like spam filtering, malware detection and network intrusion detection. Presence of adversary in these applications intends to modify distribution of test data to mislead detection system.

Several studies used game theory approaches to model the presence of an adversary during the learning phase. In [8], authors introduced an adversarial classifier as a reverse engineering learning problem to address the inherent adversarial nature of spam filtering application. In [9] an adversary model is anticipated during the learning process. According to this model, the adversary deletes and corrupts a subset of features during the classification phase. Therefore, the presence of the adversary destroyed stationarity of feature distributions in training and test time. In this approach learner's and attacker's loss functions are antagonistic. There are some studies that focused on non-antagonistic learner's and attacker's loss functions. Brückner et al. [10] modeled the interaction between classifier learner and adversary as a Stackelberg competition in which the classifier plays the role of the leader and the adversary may react according to the leader's move. They solved an optimization problem in order to explore the solution of this game. In [11] the interaction between a learner and an adversary is modeled as a static game. They assumed some conditions to find a unique Nash equilibrium and derive algorithms that find the equilibria of the prediction model.

In general, game-theoretic methodologies try to establish secure learning and develop a suitable countermeasure. However, these approaches relax a lot of constraints, because the compatibility with realistic constraint is too complex to be incorporated into existing game-theoretic approaches [12].

In [13] a proper security evaluation framework named as proactive arms race is introduced. The aim of this framework is to provide a more robust and secure detection system. The activities of a classifier designer in the proactive arms race are illustrated in Figure 1; these activities are described as the following steps:

- First step: identify potential attacks;
- Second step: simulate the identified attacks;
- Third step: evaluate the impact of simulated attacks;
- Fourth step: if the simulated attacks have a significant effect on learning systems, then develop a suitable countermeasure.

The recent studies in the proactive arms race explored the aforementioned activities of a classifier designer. In [12] Biggio et al. investigated the vulnerabilities of classification algorithms by devising evasion attack at test time. This attack manipulates malicious test samples to misclassify and evade detection system. The attack strategy is inspired by [14] which is depended on discriminative directions technique. In the recent studies [15], [16], [17], [18] the effect of using reduced feature set against adversarial attacks is investigated; the adversarial feature selection aims to improve robustness of learning system under attack. Evaluation results confirm that traditional feature selection methods are not adversary aware techniques; therefore, they may get worsened under attack. Zhange et al. proposed an adversarial-aware feature selection against evasion attack [15] by addressing classifier security in feature selection process. They demonstrated that their approach outperforms the traditional feature selection techniques. In this paper, we employ the proactive arms race framework for security evaluation of SVM classifier as a target system on the application of PDF malware detection. Since standard SVM is not an adversary-aware detection system, its performance degrades significantly in the adversarial environment. In this work, we proposed an adversary-aware learning model that is robust under evasion attack at test time.

III. Background on Evasion Attack

There are different types of attacks against machine learning systems [19]. According to evasion attack that is considered in this paper, distribution of malicious test samples are changed to evade detection system and increase false negative rate of system. This is the most typical attack in security application. In the following, we define the notations to describe the adversary's strategy. Given $f\colon \mathcal{X} \rightarrow \mathcal{Y}$ as a two-class classification algorithm that assigns samples in the feature space $x \in \mathcal{X}$ to a class label $y \in \mathcal{Y} = \{-1, 1\}$, where -1, $+1$ indicate the legitimate, malicious class, respectively. The classifier f is trained on data $\mathcal{D}$ from distribution $P(\mathcal{X}; \mathcal{Y})$. The classifier predicts the class labels using a threshold on continuous discriminate function $g\colon \mathcal{X} \rightarrow \mathcal{R}$. The predicted labels are denoted by $y^c = f(x)$, where y shows true labels. In addition, a sample with $g(x) < 0$ is labeled as negative class, i.e. $f(x) = -1$. $g(x) > 0$ indicates a sample of positive class with $f(x) = +1$.

Adversary's strategy for devising an evasion attack can be implemented through different optimization problems [12], [11], [8]. We leverage the optimization problem

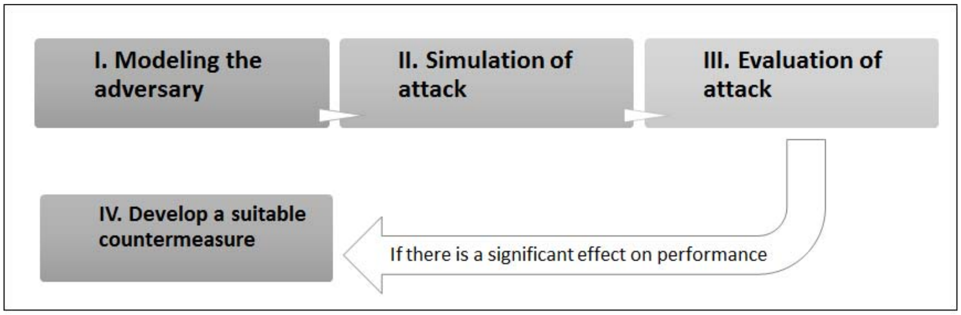

I. Modeling the adversary
II. Simulation of attack
III. Evaluation of attack
IV. Develop a suitable countermeasure
If there is a significant effect on performance

Figure 1: Activities of a classifier designer in the proactive arms race framework.

that has been presented in [12]. In this optimization problem, for any target malicious sample x_0, an adversary finds a sample x_0^* by minimizing $g(x)$, subject to a constraint on its distance from x_0:

$$x_0^* = \operatorname*{argmin}_x \; g(x) \quad s.t. \;\; d(x, x^0) \leq d_{max} \tag{1}$$

Maximum distance, d_{max}, determines the maximum allowable modification of a sample under attack. For instance, in the spam filtering the maximum distance is the maximum number of words which are allowed to be changed by an adversary; with this constraint the semantic of a spam message is preserved. Thus, this distance can be beneficial in controlling how samples could be changed and preserve their previous nature. In addition, d_{max} presents the strength of the attack, and gives the adversary freedom to modify malicious samples. Higher d_{max} results in higher changes in the features' value. Defining a suitable d_{max} is application dependent.

IV. Learning a Secure Classifier

Studies in the adversarial learning focus on two main steps of proactive arms race that are: modeling the identified attacks; and then developing a proper countermeasure to learn a secure classifier. In this section, first we present an evasion attack against classifier with differentiable discriminant functions, like SVM classifier, by inspiring from [12]; after that, we propose a secure learning model against the evasion attack on the application of malware detection in PDF files. PDF files impose a constraint during devising an evasion attack. According to PDF file structure, it is more manageable to insert an object into PDF file and preserve its structure than removing an object from PDF file; with this limitation, during an attack features' value of a target PDF are allowed to be only incremented. Due to this constraint, we consider two assumptions during devising an attack and developing a countermeasure against the modeled attack on PDF data.

- ASSUMPTION 1. Feature addition constraint: This constraint implies the limitation of PDF file structures; generally, this constraint should be applied in the applications in which features' value of malicious sample can only be incremented during evasion.
- ASSUMPTION 2. Both classifier designer and adversary are aware of feature addition constraint.

A. Modeling Evasion Attack

Algorithm 1 presents an implementation of evasion attack according to Equation 1 via gradient-based optimization algorithm. This algorithm assumes that the discriminate function of classifier is differentiable. The algorithm manipulates a target malicious sample based on adversary's knowledge in relation with discriminant function of target classifier, and changes it such that it can mislead the target classifier and increase false negative of the detection system. In this algorithm, x_0 denotes a malicious sample which is targeted by the adversary, and d_{max} is the aforementioned attack parameter. Higher d_{max} results in a severe attack. $g(x)$ is the discriminant function of target classifier. Feature addition constraint is accounted by setting the constraint, $x_k < x_{k-1}$, in line 5 where constraints are defined.

B. SVMPW: soft-margin SVM with Positive Weight

Given a training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i \in \mathbf{R}^n$ and $y_i \in \{-1, 1\}$, the objective is to learn a linear scoring function $g(x) = w^T x + b$ with model parameter $w \in \mathbf{R}^m$ and intercept $b \in \mathbf{R}$; while the proposed model is robust against evasion attack. With minimizing the regularized training error, the model parameter is estimated as Equation 2,

$$E(w, b) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, g(x_i)) + \alpha \; R(w) \tag{2}$$

where L is a loss function that measures model misclassification and R is a regularization term that penalizes model complexity; $\alpha > 0$ is a non-negative hyperparameter.

Algorithm 1 **Evasion Attack**

!ph

Input: target sample,x_0; the step size,η; maximum allowable modification,d_{max}.

Output: the optimal attacked sample, x^*.

```
$k \leftarrow 0$
repeat
    $k \leftarrow k+1$
    $x_k \leftarrow x_{k-1} - \eta \ \nabla g(x)$
    if $d(x_k, x_{k-1}) > d_{max} \ \ or \ \ x_k \ < \ x_{k-1}$ then
        Project $x_k$ onto the boundary of the feasible region constraint.
    end if
until $g(x)_k - g(x)_{k-1} < \epsilon$
return $x^* = x_k$
```

Algorithm 2 **SVMPW**

Input: the step size, η; $\epsilon > 0$.

Output: optimal model parameters, w and b.

```
$k \leftarrow 0.$
$w_0, b_0 \leftarrow$ initial values
repeat
    $k \leftarrow k+1$
    $w_k \leftarrow w_{k-1} - \eta(\alpha \frac{\partial R(w)}{\partial w} + \frac{\partial L(w^T x_i + b, y_i)}{\partial w})$
    $b_k \leftarrow b_{k-1} + \eta \frac{\partial L(w^T x_i + b, y_i)}{\partial b}$
    if $w_k < 0$ then
        Project $w_k$ onto the boundary of the feasible region constraint.
    end if
until $E(w,b)_k - E(w,b)_{k-1} < \epsilon$
return $w^* = w_k, b^* = b_k$
```

We define a soft margin Support Vector Machine scoring model using Hinge loss function. Hinge loss function $\sum_i max(0, 1 - y_i(w^T x_i + b)$ in the soft-margin SVM penalizes misclassifications. In addition, l_2-norm is chosen for the regularization term R as a common choice. After making substitution Equation 3 is computed as

$$E(w,b) = \frac{1}{n} \sum_{i=1}^{n} max(0, 1 - y_i(w^T x_i + b) + \alpha \ \frac{1}{2} \sum_{i=1}^{n} w_i^2 \tag{3}$$

As a proper countermeasure to the modeled evasion attack, the classifier designer should explore a solution to preserve its robustness under this attack. The classifier designer's strategy is to improve robustness of learning system by estimating appropriate model parameters. As mentioned, Adversary modifies features' value of the target sample as follow,

$$x_k \leftarrow x_{k-1} - \eta \ (\nabla g(x) = w)$$

The classifier designer estimates the parameter model w with only positive or zero values. Learning positive weight along with feature addition limitation results in a more secure detection model; considering the feature addition constraint, adversary has no chance to find a descent path in order to update features value of a target sample. Algorithm 2 solves the following optimization problem that learns a soft margin SVM with positive weight, SVMPW, via gradient descent procedure.

$$w^*, \ b^* = \underset{w,b}{\text{argmin}} \ \ E(w,b) \qquad s.t. \quad w \geq 0$$

The outputs of this algorithm are the optimal model parameter w and intercept b, while vector w takes only positive or zero values.

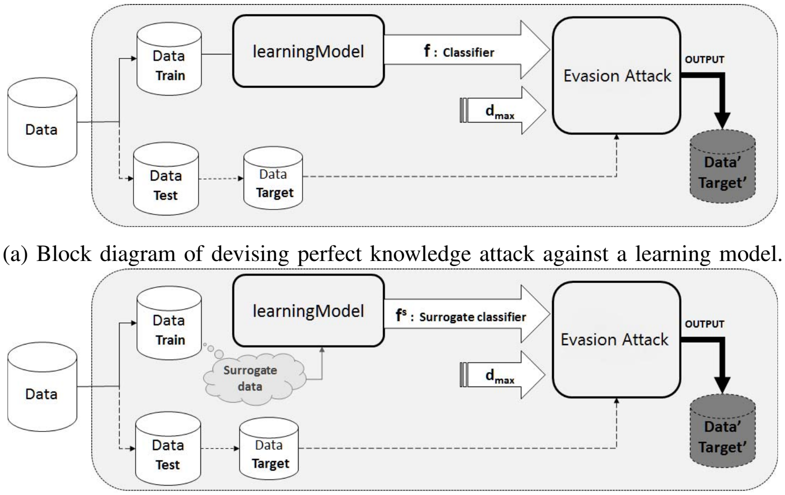
V. Security Evaluation of Learning System

To evaluate the robustness of SVMPW and standard SVM with RBF and linear kernels against evasion attack at test time, we leverage the proactive arms race framework; in addition, we employ Algorithm 1 in order to apply evasion attack. In devising an evasion attack, the adversary's knowledge range could be from no information to complete information. In the worst case an attacker has a perfect knowledge (PK) of a target system. In the PK-attack an adversary has comprehensive information about the feature space, the type of classifier, and the trained model. Although, it is an unrealistic assumption, it helps a classifier designer to determine a lower bound performance of a learning system under attack. A more realistic assumption is referred to as limited knowledge (LK) attack that adversary 's knowledge is limited to general information about type of classifier and feature space not the trained model. However, the adversary can collect a surrogate data set $\mathcal{D}_{surrogate}$ of the same distribution $P(x;y)$ from which $\mathcal{D}$ was drawn. This data could be sampled by the adversary in different ways. For instance, a spammer can collect legitimate and spam emails from an alternative source. In addition, the adversary needs to approximate the discriminant function $g(x)^s$ of the surrogate classifier learned on $\mathcal{D}_{surrogate}$. Figure 2 presents block diagrams of devising perfect knowledge versus limited knowledge attack.

In Algorithm 1 feasible region boundary is defined subject to these constraints:

- $0 \leq x_k \leq 1$, features' value take positive value in the range of $[0, 1]$,
- $x_k > x_{k-1}$, feature addition constraint,
- $\|(x_k - x_{k-1})\| > d_{max}$, l_1-norm distance between the modified sample and its original point should be lower than d_{max}. According to related work, we assume that the adversary prefers to significantly manipulate as few features as possible [15], [12]. Therefore, in this condition l_1-norm is chosen, because it converges to the sparse solution in comparison to $l_2 - norm$.

The minimize function gets discriminate function of the



(a) Block diagram of devising perfect knowledge attack against a learning model.

(b) Block diagram of devising limited knowledge attack against a learning model.

Figure 2: Block diagram of devising evasion attack against a learning model. The output is the modified target data with different distribution.

detection system as an objective function. For instance, discriminant function of standard SVM classifier and its gradient are defined as Equations 4 and 5, respectively.

$$g(x) = \sum_i \alpha_i \; y_i \; k(x, x_i) + b \tag{4}$$

$$\nabla \; g(x) = \sum_i \alpha_i \; y_i \; \nabla \; k(x, x_i) \tag{5}$$

where k is the kernel function and $\nabla k(x, x_i)$ is gradient of kernel function; in case of RBF kernel, $k(x, x_i) = exp\{-\gamma \parallel x - x_i \parallel^2\}$, is computed using Equation 6.

$$\nabla k(x, x_i) = -2\gamma \; exp\{-\gamma \parallel x - x_i \parallel^2\}(x - x_i) \tag{6}$$

Finally, the output of $EvasionAttack$ is the attacked data that are mostly misclassified by the target classifier. The classifier designer uses the modeled evasion attack to evaluate the performance degradation of the learning system in presence of the adversary.

VI. Experimentation and Evaluation

In this section, we apply evasion attack on a well-known application in adversarial learning that is malware detection in PDF file format. We investigate robustness of SVMPW, SVM with linear and RBF kernels under PK and LK evasion attacks at test time where SVMPW is an adversary aware learning model. We experimentally show how adversary threatens assets of a precise learning model like SVM.

A. Malware Detection in PDF Files

We consider the PDF malware dataset including 5591 legitimate and 5993 malicious PDFs that has been explored in [15]. Each feature value in this dataset represents the occurrence of a special keyword; the first 1,000 PDFs are considered to generate 114 distinct keywords as a feature set. In addition, the maximum value of each feature is bounded to 100. After that all features are normalized to the range of [0, 1] by dividing them by 100.

In the experiment, the maximum number of added keywords, d_{max}, controls the attack power. We consider the maximum number of 60 added keyword for modification, i.e. $d_{max} \in \{0, 60\}$. The statistical evaluation is applied by randomly splitting samples into two portions named train and test splits. Each split contains both malicious and legitimate PDFs. Fifty percent of the samples were employed to the training split and the remaining portion used for testing split in order to evaluate classification performance on the test time. The above process is repeated five times to average final result of the classifier. Moreover, we run a 5-fold cross validation on the training portion to estimate the classifier's parameters that mostly yielding C = 100 and gamma = 0.1. True Positive (TP) rate indicates the portion of malicious samples that classified correctly. We used average TP rate at 10% false positive operating point to evaluate performance of classifier because in the frequency application typically performance is reported at low false positive rate [20], [13].

1) Evaluation Results: In this section, first we evaluate standard performance of target learning models (SVMPW and SVM) then we assess robustness of them under evasion attack at test time. Generally, performance evaluation of

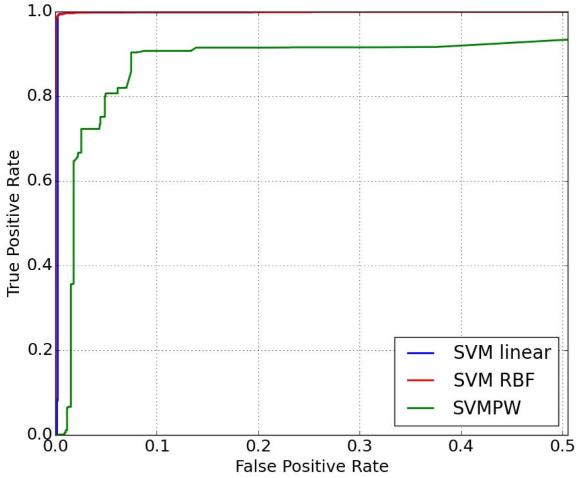


Figure 3: ROC curves of target classifiers for one of the five runs performed on PDF dataset. The plot focuses on FPR = 0.5.

learning system according to both generalization capability and robustness under attack is a open research study in the adversarial classification problem. There is a trade off between learning a more accurate classifier and designing a more secure learning model. The role of classifier designer is to identify potential vulnerability of the model, and then developing a suitable countermeasure in the response of attack. Figure 3 depicts standard performance of SVMPW and SVM with linear and RBF kernels. This figure indicates that the standard SVM precisely detects test samples and achieves high detection rate on test time while SVMPW exhibits a performance loss in comparison to standard SVM. SVMPW enforces the linear scoring model to estimate model parameter W with positive values in order to provide a more resistant model against attack. To compare security evaluation of the aforementioned classifiers, we simulate the effect of adversary using Algorithm 1 on test data and provide a test set that is affected by adversary; the test set follows a different distribution that is not the same as train data.

Figure 4 depicts changes in the average value and standard deviation of True Positive (TP) at 10% False Positive (FP) for the standard SVM with linear and RBF kernels, and SVMPW in terms of maximum number of added keywords between 0 and 60. In Figure 4, the average TP rate in the maximum number of modified words at zero, i.e. $d_{max} = 0$, shows performance of classifiers before the attack; that is the standard performance of learning system where train and test data have the same distribution. As Figure 3 confirms average TP at this point indicates that SVMPW has a performance loss in case of no attack. That is the side effect of estimating positive weight. Although detection of SVMPW is not as precise as standard SVM, its TP curve is thoroughly robust against increasing attack power. Increasing the number of added keywords shows a more powerful effect of adversary. Figure 4 demonstrates that TP curves of standard SVM with linear and RBF kernels are extremely downward with increasing attack power, i.e. d_{max}. To investigate robustness of classifiers, we devise PK and LK attacks. The solid curves in the first row of Figure 4 show performance of learning systems under PK attack. Furthermore, the dash curves in second row indicate the effect of LK attack.

To model a LK attack we provide a smaller training data as surrogate data set which contains only 500 samples, then we use the learned surrogate classifier to devise LK attack against the real trained model. The experimental evaluation approves that PK attack has a severe effect on the classifier's performance, because in PK attack the adversary has the comprehensive information about the target classifier. Hence, LK attack evades the target classifier with a lower probability in comparison to PK attack. Since SVMPW represents a robust model against evasion attack, the different attack types have no impact on this classifier.

In both attack types, performance degradation of standard SVM with RBF kernel is lower than linear one. Finally, experimental results acknowledge that SVMPW proposes a secure and robust learning model against evasion attempt of adversary at test time. As a matter of fact, choosing a system that is more robust under attack is more reasonable than a system which is more accurate according to the standard evaluation of performance; because performance degradation of the second system is significantly higher than the first one under attack, as it is shown in Figure 4.

VII. Concluding Remarks and Future Work

The challenge between attackers and classifier designers leads to an arms race; typically, this is a reactive arms race. In a reactive arms race an attacker analyzes a learning system and devises an attack to evade it. The classifier designer reacts by appropriate countermeasures like: analyzing the new samples and re-training the model by the new collected samples. In order to provide a secure learning system, the classifier designer should proactively evaluate its security by simulating a proactive arms race. In the proactive arms race, the classifier designer evaluates the classifier's security during its design phase by identifying potential attacks, modeling the identified attacks, and evaluating the impact of the modeled attacks; finally, if the modeled attacks have a significant effect on the learning system, the designer develops a suitable countermeasure. In this paper, we leverage proactive arms race framework and presented a secure learning model against an evasion attack named as SVMPW that is a soft margin SVM with

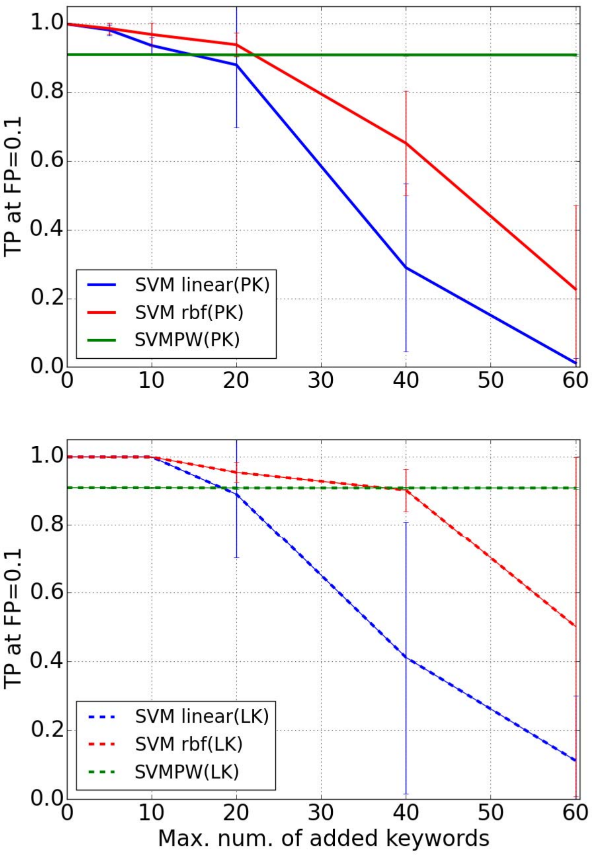


Figure 4: The average value and standard deviation of True Positive (TP) at 10% False Positive (FP) for standard SVM with linear and RBF kernels and SVMPW. Average TP value ±half standard deviation are shown with error bars.

Positive Weight parameter. Evasion attack is a typical attack caused by adversary in the adversarial environment. This attack aims to modify the distribution of malicious samples of test data to resemble legitimate samples. The experimental results confirm SVMPW effectively improves robustness of the detection system against the evasion attack at test time and outperforms the standard SVM with RBF and linear kernels. As a future work we intend to modify SVMPW in such a way to increase its accuracy in the absence of attack and improve its standard performance like standard SVM. Moreover, we plan to propose a non-linear and secure learning model against the evasion attack at test time.

References

[1] Ali Feizollah, Nor Badrul Anuar, Rosli Salleh, Fairuz Amalina, Rauf Ridzuan Maarof, and Shahaboddin Shamshirband. A study of machine learning classifiers for anomaly-based mobile botnet detection. *Malaysian Journal of Computer Science*, 26(4), 2014.

[2] Abdullah Gani. Identifying false alarm for network intrusion detection system using hybrid data mining and decision tree. *Malaysian journal of computer science*, 21(2), 2008.

[3] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pages 16–25. ACM, 2006.

[4] Davide Maiorca, Igino Corona, and Giorgio Giacinto. Looking at the bag is not enough to find the bomb: an evasion of structural methods for malicious pdf files detection. In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, pages 119–130. ACM, 2013.

[5] Davide Maiorca, Giorgio Giacinto, and Igino Corona. A pattern recognition system for malicious pdf files detection. In *Machine Learning and Data Mining in Pattern Recognition*, pages 510–524. Springer, 2012.

[6] Charles Smutz and Angelos Stavrou. Malicious pdf detection using metadata and structural features. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 239–248. ACM, 2012.

[7] Nedim Šrndic and Pavel Laskov. Detection of malicious pdf files based on hierarchical document structure. In *Proceedings of the 20th Annual Network & Distributed System Security Symposium*, 2013.

[8] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647. ACM, 2005.

[9] Ofer Dekel, Ohad Shamir, and Lin Xiao. Learning to classify with missing and corrupted features. *Machine learning*, 81(2):149–178, 2010.

[10] Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555. ACM, 2011.

[11] Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, 13(1):2617–2654, 2012.

[12] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer, 2013.

[13] Battista Biggio, Giorgio Fumera, and Fabio Roli. Security evaluation of pattern classifiers under attack. *Knowledge and Data Engineering, IEEE Transactions on*, 26(4):984–996, 2014.

[14] Polina Golland. Discriminative direction for kernel classifiers. In *Advances in neural information processing systems*, pages 745–752, 2001.

[15] Fei Zhang, Patrick PK Chan, Battista Biggio, Daniel S Yeung, and Fabio Roli. Adversarial feature selection against evasion attacks. 2015.

[16] Fei Wang, Wei Liu, and Sanjay Chawla. On sparse feature attacks in adversarial learning. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 1013–1018. IEEE, 2014.

[17] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1689–1698, 2015.

[18] Bo Li and Yevgeniy Vorobeychik. Feature cross-substitution in adversarial classification. In *Advances in Neural Information Processing Systems*, pages 2087–2095, 2014.

[19] Marco Barreno, Blaine Nelson, Anthony D Joseph, and JD Tygar. The security of machine learning. *Machine Learning*, 81(2):121–148, 2010.

[20] Aleksander Kołcz and Choon Hui Teo. Feature weighting for improved classifier robustness. In *CEAS09: sixth conference on email and anti-spam*, 2009.