

Adversarial Feature Selection against Evasion Attacks

Fei Zhang, *Student Member, IEEE*, Patrick P.K. Chan, *Member, IEEE*, Battista Biggio, *Member, IEEE*,
Daniel S. Yeung, *Fellow, IEEE*, Fabio Roli, *Fellow, IEEE*

Abstract—Pattern recognition and machine learning techniques have been increasingly adopted in adversarial settings such as spam, intrusion and malware detection, although their security against well-crafted attacks that aim to evade detection by manipulating data at test time has not yet been thoroughly assessed. While previous work has been mainly focused on devising adversary-aware classification algorithms to counter evasion attempts, only few authors have considered the impact of using reduced feature sets on classifier security against the same attacks. An interesting, preliminary result is that classifier security to evasion may be even worsened by the application of feature selection. In this paper, we provide a more detailed investigation of this aspect, shedding some light on the security properties of feature selection against evasion attacks. Inspired by previous work on adversary-aware classifiers, we propose a novel adversary-aware feature selection model that can improve classifier security against evasion attacks, by incorporating specific assumptions on the adversary’s data manipulation strategy. We focus on an efficient, wrapper-based implementation of our approach, and experimentally validate its soundness on different application examples, including spam and malware detection.

Index Terms—Adversarial Learning, Feature Selection, Classifier Security, Evasion Attacks, Spam Filtering, Malware Detection

I. INTRODUCTION

MACHINE learning has been widely used in security-related tasks such as biometric identity recognition, malware and network intrusion detection, and spam filtering, to discriminate between malicious and legitimate samples (e.g., impostors and genuine users in biometric recognition systems; spam and ham emails in spam filtering) [1]–[7]. However, these problems are particularly challenging for machine learning algorithms due to the presence of intelligent and adaptive adversaries who can carefully manipulate the input data to downgrade the performance of the detection system, violating the underlying assumption of data stationarity, *i.e.*, that training and test data follow the same (although typically unknown) distribution [8]–[12]. This has raised the issue of understanding whether and how machine learning can be

securely applied in adversarial settings, including its vulnerability assessment against different, potential attacks [13].

In one relevant attack scenario, referred to as *evasion attack*, attackers attempt to evade a deployed system at test time by manipulating the attack samples. For instance, spam, malware and network intrusion detection can be evaded by obfuscating respectively the content of spam emails (e.g., by misspelling bad words like “cheap” as “che@p”), and the exploitation code embedded in malware samples and network packets [5], [7], [10]–[12], [14]–[19]. Previous work on evasion attacks has mainly investigated the vulnerability of supervised [7], [11], [20] and unsupervised learning techniques [21], [22] in different applications, including spam filtering [5], [23], [24], intrusion detection [25] and biometric recognition [1], [2]. Few studies have instead addressed the problem of training data *poisoning* to mislead classifier learning [6], [26]–[29].

Research in adversarial learning has not only been addressing the problem of evaluating security of current learning algorithms to carefully-targeted attacks, but also that of devising learning algorithms with improved security. To counter evasion attacks, explicit knowledge of different kinds of adversarial data manipulation has been incorporated into learning algorithms, *e.g.*, using game-theoretical [8], [12], [20], [30] or probabilistic models of the hypothesized attack strategy [2], [9]. Multiple classifier systems, which have been originally proposed to improve classification accuracy through the combination of weaker classifiers, have also been exploited to the same end [15], [16], [31]. Countermeasures to poisoning attacks have also been proposed, based on data sanitization (*i.e.*, a form of outlier detection) [6], [32], multiple classifier systems [33], and robust statistics [26].

While previous work has been mainly focused on devising *secure* classification algorithms against evasion and poisoning attempts, only few authors have considered the impact of using reduced feature sets on classifier security against the same attacks. An interesting result is that classifier security to evasion may be even worsened by the application of feature selection, if adversary-aware feature selection procedures are not considered [11], [16], [34]–[36]. In particular, it has been shown that using reduced feature sets may require an attacker to manipulate less features to reach a comparable probability of evading detection, *i.e.*, given the same amount of manipulations to the attack samples, the probability of evading detection can be higher; *e.g.*, in spam filtering, using a smaller dictionary of selected words (*i.e.*, features) may not affect accuracy in the absence of attack, but it may significantly worsen classifier *security*, *i.e.*, its performance under attack.

F. Zhang, P. P.K. Chan, D. S. Yeung are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

F. Zhang: e-mail zjfei87@gmail.com, phone +86 20 3938 0285 (3415)

P. P. K. Chan (corresponding author): e-mail patrickchan@ieee.org, phone +86 20 3938 0285 (3415)

D. S. Yeung: e-mail danyeung@ieee.org, phone +86 20 3938 0285 (3304)

B. Biggio and F. Roli are with the Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d’Armi, 09123 Cagliari, Italy

B. Biggio: e-mail battista.biggio@diee.unica.it, phone +39 070 675 5776

F. Roli: e-mail roli@diee.unica.it, phone +39 070 675 5779

The above result has questioned the suitability of feature selection approaches for adversarial settings, *i.e.*, whether and to what extent such techniques can be applied without affecting classifier security against evasion (and poisoning) attacks. To our knowledge, this issue has only been recently investigated, despite the relevance of feature selection in classification tasks. Selecting a relevant subset of features may indeed not only improve classifier generalization, but it may also significantly reduce computational complexity and allow for a better data understanding [37], [38]. Therefore, understanding whether these advantages can be exploited without compromising system security in security-sensitive tasks (where reducing computational complexity is of particular interest due to the massive amount of data to be processed in real time) can be considered a relevant, open research issue.

In this paper, we present a systematic security evaluation of classification algorithms exploiting reduced feature sets to gain a better understanding of how feature selection may affect their security against evasion attacks. We further propose a feature selection model that allows us to improve classifier security against the same attacks, while using a significantly reduced feature representation (Sect. III). The underlying idea of our approach is to select features not only based on the generalization capability of the resulting classifier in the absence of attack (as in traditional feature selection methods), but also depending on its security against adversarial data manipulation. We model classifier security as a regularization term to be optimized together with the estimated classifier's generalization capability during the feature selection process. The proposed model is implemented as a wrapper-based feature selection approach, suitable for linear and non-linear classifiers (with differentiable discriminant functions), and for discrete- and real-valued feature spaces. We exploit the well-known forward selection and backward elimination algorithms to implement the proposed approach. Its effectiveness against attacks that assume different levels of knowledge of the attacked system (discussed in Sect. IV) is experimentally evaluated on different application examples, including spam and PDF malware detection (Sect. V). We finally discuss contributions, limitations and future work (Sect. VI).

II. BACKGROUND

In this section we revise some useful concepts that will be exploited in the rest of the paper, also introducing our notation. We start by describing previously-proposed measures of *classifier security* (or robustness) against evasion attacks. We then discuss traditional feature selection methods, and their *stability* to non-adversarial perturbations.

A. Adversarial Attacks and Classifier Security to Evasion

An implicit assumption behind traditional machine learning and pattern recognition algorithms is that training and test data are drawn from the same, possibly unknown, distribution. This assumption is however likely to be violated in adversarial settings, since attackers may carefully manipulate the input data to downgrade the system's performance [8], [10]–[12], [39]. A taxonomy of potential attacks against machine learning

has been defined in [10], [13], [39]. It categorizes attacks according to three axes: the attack influence, the kind of security violation, and the attack specificity. The **attack influence** can be either *causative* or *exploratory*. A causative (or *poisoning*) attack alters the training data to mislead subsequent classification of test samples [29], [40], while an exploratory (or *evasion*) attack directly manipulates test samples to cause misclassifications [10], [11], [13], [39]. Depending on the kind of **security violation**, an attack may compromise a system's *availability*, *integrity*, or *privacy*: availability attacks aim to downgrade the overall system's accuracy, causing a denial of service; integrity attacks, instead, only aim to have malicious samples misclassified as legitimate; and privacy attacks aim to retrieve some protected or sensitive information from the system (*e.g.*, the clients' biometric templates in biometric recognition systems). The **attack specificity** defines whether the attack aims to change the classifier decision on a *targeted* set of samples, or on an *indiscriminate* fashion (*e.g.*, if the goal is to have *any* malicious sample misclassified as legitimate). This taxonomy has been subsequently extended in [11], [19] by making more detailed assumptions on the attacker's goal, knowledge of the targeted system, and capability of manipulating the input data, to allow one to formally define an *optimal* attack strategy. Notably, in [21], a similar model of attacker has been proposed to categorize attacks against unsupervised learning algorithms (*i.e.*, clustering).

According to the aforementioned taxonomy, the evasion attack considered in this paper can be regarded as an exploratory integrity attack, in which malicious test samples are manipulated to evade detection by a classifier trained on untainted data. This is indeed one of the most common attacks in security-related tasks like spam filtering [5], [23], [24], network intrusion detection [25], and biometric recognition [1], [2]. Optimal evasion has been formulated according to slightly different optimization problems [11], [12], [16], [19], [41], [42]. In general, the rationale behind all of the proposed attacks is to find a sample $\mathbf{x}^* \in \mathcal{X}$ that evades detection by *minimally* manipulating the initial attack $\mathbf{x} \in \mathcal{X}$, where the amount of manipulations is characterized by a suitable distance function in feature space. For instance, in [41], [42] optimal evasion is formulated as:

$$\begin{aligned} \mathbf{x}^* &= \arg \min_{\mathbf{x}'} c(\mathbf{x}', \mathbf{x}) \\ \text{s.t.} \quad &g(\mathbf{x}') < 0, \end{aligned} \quad (1)$$

where $c(\mathbf{x}', \mathbf{x})$, with $c : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, is the distance of the manipulated sample from the initial attack, and \mathbf{x}' is classified as legitimate if the classifier's discriminant function $g : \mathcal{X} \mapsto \mathbb{R}$ evaluated at \mathbf{x}' is negative. Without loss of generality, this amounts to assuming that the classification function $f : \mathcal{X} \mapsto \mathcal{Y}$, with $\mathcal{Y} = \{-1, +1\}$ can be generally written as $f(\mathbf{x}') = \text{sign}(g(\mathbf{x}'))$, being -1 and $+1$ the legitimate and the malicious class, respectively, and $\text{sign}(a) = +1$ (-1) if $a \geq 0$ ($a < 0$). A typical choice of distance function for Boolean features is the Hamming distance, which amounts to counting the number of feature values that are changed from \mathbf{x} to \mathbf{x}' by the attack. In spam filtering, this often corresponds to the number of modified words in each spam, having indeed

a meaningful interpretation [11], [12], [15], [16], [41], [42].

Based on the above definition of optimal evasion, in [34], [43] the authors have proposed a measure of classifier security against evasion attacks, called *hardness of evasion*, to show that multiple classifier systems that average linear classifiers can be exploited to improve robustness to evasion. It is simply defined as the expected value of $c(\mathbf{x}^*, \mathbf{x})$ computed over all attack samples. In the case of Boolean features, it amounts to computing the average minimum number of features that have to be modified in a malicious sample (e.g., the minimum number of words to be modified in a spam email) to evade detection. A different measure of classifier security to evasion, called *weight evenness*, has been later proposed in [15], [16]. It is based upon the rationale that a robust classifier should not change its decision on a sample if only a small subset of feature values are modified. For linear classifiers, this can be easily quantified by measuring whether the classifier's weights are evenly distributed among features, since more evenly-distributed feature weights should require the adversary to manipulate a higher number of features to evade detection. Accordingly, weight evenness has been defined as:

$$E = \frac{2}{d-1} \left[d - \sum_{k=1}^d \left(\frac{\sum_{i=1}^k |w_{(i)}|}{\sum_{j=1}^d |w_{(j)}|} \right) \right] \in [0, 1], \quad (3)$$

being $|w_{(1)}| \geq |w_{(2)}| \geq \dots \geq |w_{(d)}|$ the absolute values of the classifier's weights sorted in descending order, and d the number of features. Higher values of E clearly correspond to even weight distributions. This measure has also been exploited to improve the robustness of support vector machines (SVMs) and multiple classifier systems against evasion attacks [15], [16], [20].

B. Feature Selection, Robustness, and Stability

Feature selection is an important preprocessing step in pattern recognition [38], [44], [45]. It has been widely used in bioinformatics [37], [46], image steganalysis [47], [48], network intrusion detection [49], camera source model identification [50] and spam detection [16], [51]. Its goal is to choose a relevant subset of features not only to improve a classifier's generalization capability when only few training samples are available, but, most importantly, to reduce time and space complexity [38]. Another advantage is that data understanding and visualization are also facilitated after removing irrelevant or redundant features [46].

Feature selection methods can be divided into three categories according to their interaction with classification algorithms [38], [44], [52]. *Filter approaches* rank feature subsets mostly independently from the accuracy of the given classifier. For efficiency reasons, they exploit surrogate functions of the classification accuracy, based on some properties of the dataset [53]–[55], such as mutual information [56], [57]. Feature selection is instead guided by the performance of the considered classifier in *wrapper approaches*, which however require one to re-train the classification algorithm each time the feature subset is modified [51], [58]. *Embedded approaches* exploit internal information of the classifier to select features during classifier training [59], [60]. Traditional

feature selection algorithms thus optimize classification accuracy or some surrogate function with respect to the choice of the feature subset, without considering how the resulting classifier may be affected by adversarial attacks. It has indeed been shown that feature selection may even worsen classifier security to evasion: the resulting classifiers may be evaded with less modifications to the attack data [11], [16], [34]–[36].

Robust feature selection approaches have also been proposed, both to minimize the variability of feature selection results against *random* perturbations of the training data (i.e., considering different training sets drawn from the same underlying data distribution) [61], and, more recently, also to counter some kinds of adversarial data manipulations [35], [36]. As a result, the notion of 'robustness' considered in [61] is rather different from that considered in [35], [36] and in this paper. It is nevertheless of interest to understand whether methods that are more robust to evasion may also benefit from robustness to random perturbations, and vice versa.

Finally, it is worth mentioning that Robust Statistics [62], [63] may be also exploited to learn more robust feature mappings. An example is given in [10], [26], where the authors have exploited a robust version of the principal component analysis (originally proposed in [64]) to reduce the influence of poisoning attacks in the training data, yielding a more secure network traffic anomaly detector. However, to our knowledge, no work has considered thus far the problem of learning more secure classifiers against evasion attacks (i.e., manipulations of malicious samples *at test time*), by leveraging on a carefully-devised, wrapper-based feature selection approach.

III. ADVERSARIAL FEATURE SELECTION

In this section, we present our adversary-aware feature selection approach. The underlying idea is to select a feature subset that not only maximizes the generalization capability of the classifier (in the absence of attack, as in traditional feature selection), but also its security against evasion attacks. Given a d -dimensional feature space, and $m < d$ features to be selected, this criterion can be generally formalized as:

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} G(\boldsymbol{\theta}) + \lambda S(\boldsymbol{\theta}), \quad (4)$$

$$\text{s.t.} \quad \sum_{k=1}^d \theta_k = m, \quad (5)$$

where G and S respectively represent an estimate of the classifier's generalization capability and security to evasion, weighted by a trade-off parameter λ (to be chosen according to application-specific constraints, as discussed in Sect. V), $\boldsymbol{\theta} \in \{0, 1\}^d$ is a binary-valued vector representing whether each feature has been selected (1) or not (0), and $\boldsymbol{\theta}^*$ is the optimal solution.¹ Notably, the inequality constraint $\sum_{k=1}^d \theta_k \leq m$ can be alternatively considered if one aims to select the best feature subset within a maximum feature set size m .

The generalization capability $G(\boldsymbol{\theta})$ of a classifier on a feature subset $\boldsymbol{\theta}$ can be estimated according to different performance measures, depending on the given application. Provided

¹We use the same notation defined in [38], and refer to the set of selected features as $\boldsymbol{\theta}$ (although $\boldsymbol{\theta}$ is an indexing vector rather than a proper set).

that the data follows a distribution $p(\mathbf{X}, Y)$, with \mathbf{X} and Y being two random variables defined in the corresponding sets \mathcal{X} and \mathcal{Y} , and that a suitable utility function $u: \mathcal{Y} \times \mathbb{R} \mapsto \mathbb{R}$ is given, this can be formalized as:

$$G(\theta) = \mathbb{E}_{\mathbf{x}, y \sim p(\mathbf{X}, Y)} u(y, g(\mathbf{x}_\theta)), \quad (6)$$

where \mathbb{E} denotes the expectation operator, \mathbf{x}_θ is the projection of \mathbf{x} onto the set of selected features, and $g(\mathbf{x})$ is the classifier's discriminant function (see Sect. II). For instance, if $u(y, g(\mathbf{x})) = +1$ when $yg(\mathbf{x}) \geq 0$, and 0 otherwise, $G(\theta)$ corresponds to the classification accuracy. As the data distribution $p(\mathbf{X}, Y)$ is typically unknown, $G(\theta)$ can be empirically estimated from a set of available samples drawn from $p(\mathbf{X}, Y)$, as in traditional feature selection (e.g., using cross-validation).

As for the security term $S(\theta)$, we exploit the definition of minimum cost evasion given by Problem (1)-(2). Accordingly, classifier security can be defined as the *hardness of evasion* (see Sect. II-A), i.e., the average minimum number of modifications to a malicious sample to evade detection:

$$S(\theta) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{X}|Y=+1)} c(\mathbf{x}_\theta^*, \mathbf{x}_\theta), \quad (7)$$

where \mathbf{x}_θ^* is the optimal solution to Problem (1)-(2). The rationale is that more secure classifiers should require a higher number of modifications to the malicious samples to evade detection. Since, in practice, the attacker may only have limited knowledge of the system, or limited capability of manipulating the data, this should indeed yield a lower evasion rate [11], [19], [41]. The value of $S(\theta)$ can be empirically estimated from the set of available samples when $p(\mathbf{X}, Y)$ is unknown, as for $G(\theta)$, by averaging $c(\mathbf{x}_\theta^*, \mathbf{x}_\theta)$ over the set of malicious samples. Note however that this value may depend on the size of the feature subset, as it estimates an average distance measure. This may be thought as a different rescaling of the trade-off parameter λ when selecting feature subsets of different sizes. Therefore, one may rescale λ to avoid such a dependency, e.g., by dividing its value by the maximum value of $c(\mathbf{x}_\theta^*, \mathbf{x}_\theta)$ attained over the malicious samples.

In principle, the proposed criterion can be exploited for wrapper- and filter-based feature selection, provided that G and S can be reliably estimated, e.g., using surrogate measures. However, we are not aware of any technique that allows estimating classifier security to evasion without simulating attacks against the targeted classifier. We thus consider a wrapper-based implementation of our approach, leaving the investigation of filter-based implementations to future work. Two implementations of our wrapper-based adversarial feature selection, based on the popular algorithms of forward feature selection and backward feature elimination, are discussed in the next section. In the sequel, we assume that $S(\theta)$ can be estimated from the available data. We will discuss how to estimate its value by solving Problem (1)-(2) in Sect. III-B, for different choices of distance functions and classifiers.

A. Wrapper-based Adversarial Feature Selection (WAFS)

The implementation of the proposed adversarial feature selection approach is given as Algorithm 1. It is a simple variant of the popular forward selection and backward elimination wrapping algorithms, which iteratively add or delete a feature

Algorithm 1 Wrapper-based Adversarial Feature Selection, with Forward Selection (**FS**) and Backward Elimination (**BE**).

Input: $\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^n$: the training set; λ : the trade-off parameter; m : the number of selected features.

Output: $\theta \in \{0, 1\}^d$: the set of selected features.

```

1:  $\mathcal{S} \leftarrow \emptyset, \mathcal{U} \leftarrow \{1, \dots, d\};$ 
2: repeat
3:   for each feature  $k \in \mathcal{U}$  do
4:     Set  $\mathcal{F} \leftarrow \mathcal{S} \cup \{k\}$  for FS ( $\mathcal{F} \leftarrow \mathcal{U} \setminus \{k\}$  for BE);
5:     Set  $\theta = \mathbf{0}$ , and then  $\theta_j = 1$  for  $j \in \mathcal{F}$ ;
6:     Estimate  $G_k(\theta)$  and  $S_k(\theta)$  using cross-validation
       on  $\mathcal{D}_\theta = \{\mathbf{x}_\theta^i, y^i\}_{i=1}^n$  (this involves classifier training);
7:   end for
8:    $\lambda' = \lambda(\max_k S_k)^{-1}$  (i.e., rescale  $\lambda$ );
9:    $k^* = \arg \max_k (G_k(\theta) + \lambda' S_k(\theta))$ ;
10:   $\mathcal{S} \leftarrow \mathcal{S} \cup \{k^*\}, \mathcal{U} \leftarrow \mathcal{U} \setminus \{k^*\}$ ;
11: until  $|\mathcal{S}| = m$  for FS ( $|\mathcal{U}| = m$  for BE)
12: Set  $\mathcal{F} \leftarrow \mathcal{S}$  for FS ( $\mathcal{F} \leftarrow \mathcal{U}$  for BE);
13: Set  $\theta = \mathbf{0}$ , and then  $\theta_j = 1$  for  $j \in \mathcal{F}$ ;
14: Return  $\theta$ 
```

from the current candidate set, starting respectively from an empty feature set and from the full feature set [38], [44], [58]. As in traditional wrapper methods, cross-validation is exploited to estimate the classifier's generalization capability $G(\theta)$ more reliably. The *only* – although very important – difference is that our approach also evaluates the security term $S(\theta)$ to select the best candidate feature at each step.

B. Evaluating Classifier Security to Evasion

We explain here how to solve Problem (1)-(2) to estimate the classifier security term $S(\theta)$ in the objective function of Eq. (4). Problem (1)-(2) essentially amounts to finding the closest sample \mathbf{x}' to \mathbf{x} that evades detection, according to a given distance function $c(\mathbf{x}', \mathbf{x})$. In general, the problem may be solved using a black-box search approach (e.g., a genetic algorithm) that queries the classification function with different candidate samples to find the best evasion point. This approach may be however too computationally demanding, as it does not exploit specific knowledge of the objective function and of the targeted classifier. To develop more efficient algorithms, one should indeed focus on specific choices of the objective (i.e., the distance function), and of the constraints (i.e., the kind of classifier and feature representation).

As for the distance function, most of the work in adversarial learning has considered the ℓ_1 - and the ℓ_2 -norm, depending on the feature space and kind of attack [8], [19], [41], [42]; e.g., if it is more convenient for an attacker to significantly manipulate few features than slightly manipulate the majority of them, the ℓ_1 -norm should be adopted, as it promotes *sparsity*; otherwise, the ℓ_2 -norm would be a better choice.

The classification function may be linear (e.g., linear SVMs, perceptrons) or non-linear (e.g., SVMs with the radial basis function (RBF) or the polynomial kernel, neural networks). Further, it may be non-differentiable (e.g., decision trees). Previous work has addressed the evasion of linear [41] and convex-inducing classifiers, i.e., classifiers that partition the

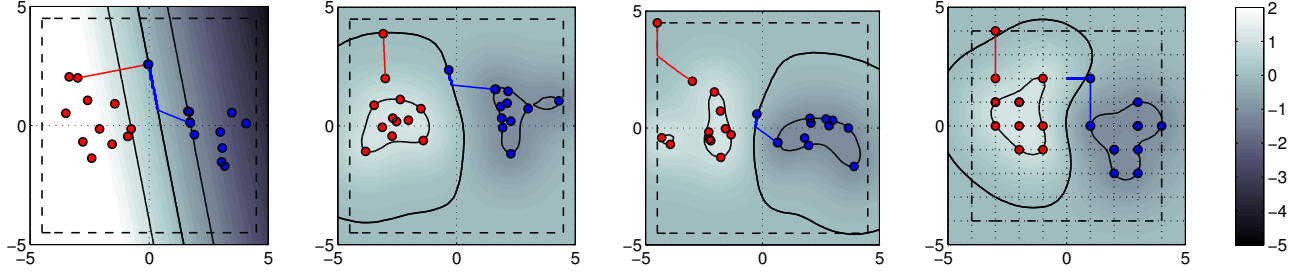


Fig. 1. Examples of descent paths obtained by Algorithm 2 to find optimal evasion points against SVM-based classifiers on a bi-dimensional dataset, for different choices of distance function, classification algorithm, and feature representation. Red and blue points represent the malicious and legitimate samples, respectively. The initial malicious sample \mathbf{x} to be modified in this example is the red point at $\mathbf{x} = [-3, 2]$. Solid black lines denote the SVM’s decision boundary (i.e., $g(\mathbf{x}) = 0$) and margin conditions (i.e., $g(\mathbf{x}) = \pm 1$). The color map represents the value of the discriminant function $g(\mathbf{x})$ at each point. Dashed black lines denote the boundaries of a box constraint. Red and blue lines denote the descent paths when the initialization point is equal to the initial malicious point ($\mathbf{x}^{(0)} = \mathbf{x}$) and to the closest point classified as legitimate, respectively. In the first and the second plot, evasion points are found by minimizing the ℓ_2 -norm $\|\mathbf{x}' - \mathbf{x}\|_2^2$, respectively against a linear SVM with regularization parameter $C = 1$, and against a nonlinear SVM with $C = 1$ using the RBF kernel with $\gamma = 0.5$, on a continuous feature space. In the third and the fourth plot, the ℓ_1 -norm $\|\mathbf{x}' - \mathbf{x}\|_1$ is minimized against the same nonlinear SVM, respectively on a continuous and on a discrete feature space, where feasible points lie at the grid intersections. The problem may exhibit multiple local minima, or have a unique solution (see, e.g., the first plot). Further, depending on the shape of the decision boundary, Algorithm 2 may not find an evasion point when initialized with $\mathbf{x}^{(0)} = \mathbf{x}$ (see, e.g., the third plot). Accordingly, the closest point to \mathbf{x} that evades detection should be eventually retained.

Algorithm 2 Evasion Attack

Input: \mathbf{x} : the malicious sample; $\mathbf{x}^{(0)}$: the initial location of the attack sample; t : the gradient step size; ϵ : a small positive constant; m : the maximum number of iterations.

Output: \mathbf{x}' : the closest evasion point to \mathbf{x} found.

```

1:  $i \leftarrow 0$ 
2: repeat
3:    $i \leftarrow i + 1$ 
4:   if  $g(\mathbf{x}^{(i)}) \geq 0$  then take a step towards the boundary
5:      $\mathbf{x}^{(i)} \leftarrow \mathbf{x}^{(i-1)} - t \nabla g(\mathbf{x}^{(i-1)})$ 
6:   else take a step to reduce the objective function
7:      $\mathbf{x}^{(i)} \leftarrow \mathbf{x}^{(i-1)} - t \nabla c(\mathbf{x}^{(i-1)}, \mathbf{x})$ 
8:   end if
9:   if  $\mathbf{x}^{(i)}$  violates other constraints (e.g., box) then
10:    Project  $\mathbf{x}^{(i)}$  onto the feasible domain
11:   end if
12: until  $c(\mathbf{x}^{(i)}, \mathbf{x}) - c(\mathbf{x}^{(i-1)}, \mathbf{x}) < \epsilon$  or  $i \geq m$ 
13: return  $\mathbf{x}' = \mathbf{x}^{(i)}$ 

```

feature space into two sets, one of which is convex [42]. It has also been recently shown that non-linear classifiers with differentiable discriminant functions can be evaded through a straightforward gradient descent-based attack [19]. We are not aware of any work related to the evasion of non-linear classifiers with non-differentiable functions. Although this may be addressed through heuristic search approaches, as mentioned at the beginning of this section, we leave a more detailed investigation of this aspect to future work. In this work we therefore consider classifiers whose discriminant function $g(\mathbf{x})$ is not necessarily linear, but differentiable. These include, for instance, SVMs with differentiable kernels (e.g., linear, RBF, polynomial) and neural networks, which have been widely used in security-related applications [11], [19].

Additional constraints to Problem (1)-(2) may be considered, depending on the specific feature representation; e.g., if features are real-valued and normalized in $[0, 1]^d$, one may consider a box constraint on \mathbf{x}' , given as $0 \leq x'_j \leq 1$, for $j = 1, \dots, d$. Further, features may take on discrete

values, making our problem harder to solve, as discussed in Sect. III-B2. In the following, we assume that feature values are continuous on a potentially compact space, such as $[0, 1]^d$.

In the easiest cases, a solution can be found analytically; e.g., if one aims to minimize $c(\mathbf{x}', \mathbf{x}) = \|\mathbf{x}' - \mathbf{x}\|_2^2$ against a linear classifier $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ (being $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ the feature weights and bias), it is easy to verify that the optimal evasion point is $\mathbf{x}' = \mathbf{x} - g(\mathbf{x}) \frac{\mathbf{w}}{\|\mathbf{w}\|_2^2}$ (cf. Fig. 1, leftmost plot).

If the discriminant function $g(\mathbf{x})$ is non-linear, the optimization problem becomes non-linear as well, and it may thus exhibit local minima. Nevertheless, a local minimum can be found by minimizing the objective function $c(\mathbf{x}', \mathbf{x})$ through gradient descent. Our idea is to take gradient steps that aim to reduce the distance of \mathbf{x}' from \mathbf{x} , while projecting the current point onto the feasible domain as soon as the constraint $g(\mathbf{x}') < 0$ is violated. In fact, following the intuition in [19], the attack point can be projected onto the non-linear, feasible domain $g(\mathbf{x}') < 0$ by minimizing $g(\mathbf{x}')$ itself through gradient descent. The detailed procedure is given as Algorithm 2. Notice however that this projection may not always be successful (see Fig. 1, third plot from the left). It may indeed happen that the attack point \mathbf{x}' reaches a flat region where the gradient of $g(\mathbf{x}')$ is null, while the point is still classified as malicious ($g(\mathbf{x}') \geq 0$). To overcome this limitation, we initialize the attack point to different locations before running the gradient descent (instead of mimicking the feature values of samples classified as legitimate, as done in [19]). In particular, we consider two initializations: one in which the attack point \mathbf{x}' is set equal to \mathbf{x} (red descent paths in Fig. 1), and the other one in which \mathbf{x}' is set equal to the closest sample classified as legitimate (blue descent paths in Fig. 1). The rationale is the following. In the former case, we start from a point which is classified as malicious, and see whether we can reach a reasonably close evasion point by following the gradient of $g(\mathbf{x}')$. In the latter case, we start from the closest point classified as legitimate, and try to get closer to \mathbf{x} while avoiding violations of the constraint $g(\mathbf{x}') < 0$. The closest point to \mathbf{x} that evades detection is eventually retained. As shown in Fig. 1, this should reasonably allow us to find

at least a good local minimum for our problem. Finally, note that the proposed algorithm quickly converges to the optimal evasion point when $g(\mathbf{x}')$ is linear too, from any of the two proposed initializations, as shown in the leftmost plot of Fig. 1.

1) *Gradients*: The gradients required to evaluate classifier security using Algorithm 2 are given below, for some distance and discriminant functions. Subgradients can be considered for non-differentiable functions, such as the ℓ_1 -norm.

Distance functions. As discussed in Sect. II, typical choices for the distance function $c(\mathbf{x}', \mathbf{x})$ in adversarial learning are the ℓ_2 - and the ℓ_1 -norm. Their gradients with respect to \mathbf{x}' can be respectively computed as $\nabla c(\mathbf{x}', \mathbf{x}) = 2(\mathbf{x}' - \mathbf{x})$, and $\nabla c(\mathbf{x}', \mathbf{x}) = \text{sign}(\mathbf{x}' - \mathbf{x})$, where $\text{sign}(\mathbf{v})$ returns a vector whose i^{th} element is 0 if $v_i = 0$, 1 (−1) if $v_i > 0$ ($v_i < 0$).

Linear classifiers. For linear discriminant functions $g(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$, with feature weights $\mathbf{w} \in \mathbb{R}^d$ and bias $b \in \mathbb{R}$, the gradient is simply given as $\nabla g(\mathbf{x}) = \mathbf{w}$.

Nonlinear SVMs. For kernelized SVMs, the discriminant function is $g(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b$, where the parameters α and b are learned during training, $k(\mathbf{x}, \mathbf{x}_i)$ is the kernel function, and $\{\mathbf{x}_i, y_i\}_{i=1}^n$ are the training samples and their labels [65]. The gradient is $\nabla g(\mathbf{x}) = \sum_i \alpha_i y_i \nabla k(\mathbf{x}, \mathbf{x}_i)$. In this case, the feasibility of our approach depends on the kernel derivative $\nabla k(\mathbf{x}, \mathbf{x}_i)$, which is computable for many numeric kernels; e.g., for the RBF kernel $k(\mathbf{x}, \mathbf{x}_i) = \exp\{-\gamma\|\mathbf{x} - \mathbf{x}_i\|^2\}$, it is $\nabla k(\mathbf{x}, \mathbf{x}_i) = -2\gamma \exp\{-\gamma\|\mathbf{x} - \mathbf{x}_i\|^2\}(\mathbf{x} - \mathbf{x}_i)$.

Although in this work we only consider kernelized SVMs as an example of nonlinear classification, our approach can be easily extended to any other nonlinear classifier with differentiable discriminant function $g(\mathbf{x})$ (see, e.g., [19], for the computation of $\nabla g(\mathbf{x})$ for neural networks).

2) *Descent in discrete spaces*: In discrete spaces, it is not possible to follow the gradient-descent direction exactly, as this may map the current sample to a set of non-admissible feature values. In fact, descent in discrete spaces amounts to finding, at each step, a feasible neighbor of the current sample that maximally decreases the objective function. This can be generally addressed using a search algorithm that queries the objective function at every point in a small neighborhood of the current sample, which may however require a large number of queries (exponential in the number of features). For classifiers with a differentiable discriminant function, the number of queries can be reduced by perturbing only a number of features which correspond to the maximum absolute values of the gradient, one at a time, in the correct direction, and eventually retaining the sample that maximally decreases the objective function. This basically amounts to exploiting the available gradient as a search heuristic, and to selecting the feasible point that best aligns with the current gradient. An example of the descent paths explored in discrete spaces by our evasion attack algorithm is given in Fig. 1 (rightmost plot).

IV. SECURITY EVALUATION

To evaluate the effectiveness of the proposed adversarial feature selection method against traditional feature selection approaches, we follow the security evaluation procedure originally proposed in [11], [19]. The underlying idea is to simulate

attacks that may be potentially incurred at test time, relying on specific assumptions on the adversary's model, in terms of his goal, knowledge of the targeted system, and capability of manipulating the input data. In the *evasion* setting, the attacker aims to evade detection by exploiting knowledge of the classification function to manipulate the malicious (test) samples. The attacker's knowledge can be either *perfect* or *limited*. In the former case, the classification algorithm is fully known to the attacker, who can then perform a worst-case attack, similarly to Problem (1)-(2). In the latter case, instead, knowledge of the *true* discriminant function $g(\mathbf{x})$ is not available. This is a more realistic case, as typically the attacker has neither access to the classifier internal parameters nor to the training data used to learn it. Nevertheless, the function $g(\mathbf{x})$ can be approximated by collecting *surrogate* data (i.e., data ideally sampled from the same distribution followed by the training data used to learn the targeted classifier), and then learning a surrogate classifier on it. As shown in [19], this can lead to very good approximations of the targeted classifier for the sake of finding suitable evasion points. Intuitively, solving Problem (1)-(2) when exploiting an approximation $\hat{g}(\mathbf{x})$ of the true discriminant function $g(\mathbf{x})$ may not be a good choice, as looking for evasion points which are only barely misclassified as legitimate by the *surrogate* classifier may lead the attacker to only rarely evade the *true* classifier. For this reason, optimal evasion has been reformulated in [19] as:

$$\min_{\mathbf{x}'} \quad \hat{g}(\mathbf{x}') , \quad (8)$$

$$\text{s.t.} \quad c(\mathbf{x}', \mathbf{x}) \leq c_{\max} , \quad (9)$$

where the constraint on $c(\mathbf{x}', \mathbf{x})$ bounds the attacker's capability by setting an upper bound on the maximum amount of modifications c_{\max} that can be made to the initial malicious sample \mathbf{x} . In this case, the malicious sample is modified to be misclassified as legitimate with the highest possible confidence (i.e., minimum value of $\hat{g}(\mathbf{x})$), under a maximum amount of modifications c_{\max} , which can be regarded as a parameter of the security evaluation procedure. In fact, by repeating the security evaluation for increasing values of c_{\max} , one can show how gracefully the performance of the true classifier decreases against attacks of increasing *strength*. The more the performance gracefully decreases, the more secure the classifier is expected to be. Examples of such curves will be shown in Sect. V. Finally, note that the aforementioned problem can be solved using an algorithm similar to Algorithm 2, in which the objective function and the constraint are exchanged.

V. APPLICATION EXAMPLES

In this section we empirically validate the proposed approach on two application examples involving spam filtering and PDF malware detection. In the former case, we compare the traditional *forward* feature selection wrapping algorithm with the corresponding implementation of our approach, using a linear SVM as the classification algorithm. In the latter case, instead, we consider traditional and adversarial *backward* feature elimination approaches, and an SVM with the RBF kernel as the wrapped classifier. We believe that these

examples can be considered a representative set of cases to assess the empirical performance of the proposed method.

A. Spam Filtering

Spam filtering is one of the most common application examples considered in adversarial machine learning [6], [10], [11], [16], [26]. In this task, the goal is often to design a *linear* classifier that discriminates between legitimate and spam emails by analyzing their textual content, exploiting the so-called *bag-of-words* feature representation, in which each binary feature denotes the presence (1) or absence (0) of a given word in an email [66]. Despite its simplicity, this kind of classifier has shown to be highly accurate, while also providing interpretable decisions. It has been therefore widely adopted in previous work [6], [10], [11], [15], [16], [23], [24], [26], and in several real anti-spam filters, like SpamAssassin and SpamBayes.² Evasion attacks against these kinds of classifier consist of manipulating the content of spam emails through bad word obfuscations (*e.g.*, misspelling spammy words like “cheap” as “che4p”) and good word insertions (*i.e.*, adding words which typically appear in legitimate emails and not in spam), which amounts to modifying the corresponding feature values from 1 to 0 and vice versa [11], [15], [16], [41].

Experimental setup. For these experiments, we considered the benchmark TREC 2007 email corpus, which consists of 25,220 legitimate and 50,199 real spam emails [67]. We represented each email as a feature vector using the tokenization method of SpamAssassin, which exploits the aforementioned bag of words feature model. To this end, we first extracted the dictionary of words (*i.e.*, features) from the first 5,000 emails (in chronological order). Then, to keep the computational complexity manageable, we reduced the feature set from more than 20,000 to 500 features, without significant loss in classification accuracy, using a supervised feature selection approach based on the information gain criterion [68].³ The linear SVM was considered as the classification algorithm. As for the performance measure $G(\theta)$, we used classification accuracy. Classifier security $S(\theta)$ was evaluated as discussed in Sect. III, using the ℓ_1 -norm as the distance function $c(\mathbf{x}', \mathbf{x})$, and Algorithm 2 for discrete spaces (Sect. III-B2). This choice of distance function amounts to counting the minimum number of words to be modified in each spam to evade detection.

We run a preliminary security evaluation to tune the trade-off parameter λ of our method, aiming to maximize the average true positive (TP) rate (*i.e.*, the fraction of correctly classified malicious samples) at the 1% false positive (FP) rate operating point (*i.e.*, when 1% of the legitimate samples are misclassified), for $c_{\max} \in [0, 20]$ (see Sect. IV). It is indeed common to evaluate system performance at low FP rates in security-related tasks [11], [15], [16]. This also allows us to compare different classifiers against evasion attacks in a fair way, as classifiers with higher FP rates may be easier to evade [11], [19]. If λ is too large, the selected features show poor generalization capability in the absence of attack (*i.e.*, when $c_{\max} = 0$), which also leads to a too

low TP rate under attack (*i.e.*, when $c_{\max} > 0$). Conversely, if λ is too small, classifier performance under attack may quickly decrease. Hence, to effectively tune λ , one should quantitatively investigate this trade-off on the available data. We assume that a maximum decrease of the TP rate in the absence of attack of 1% is tolerable, at the given FP rate. Then, the highest value of λ under this constraint can be selected, to maximize the TP rate under attack. Based on these observations, we run a 5-fold cross validation on the training set with values of $\lambda \in \{0.1, 0.5, 0.9\}$, and selected $\lambda = 0.5$.

Each experiment was repeated five times, each time by randomly selecting 2,500 samples for each class from the remaining emails in the TREC corpus. In each run, the dataset of 5,000 samples was randomly split into a training and a test set of 2,500 samples each. Then, subsets consisting of 1 to 499 features were selected according to the traditional and adversarial forward selection methods, through a 5-fold cross validation procedure on the training set, to respectively maximize the value of $G(\theta)$ and $G(\theta) + \lambda S(\theta)$ (Algorithm 1) estimated on such data. The SVM regularization parameter $C \in \{2^{-10}, 2^{-9}, \dots, 2^{10}\}$ was also selected during this process using an additional inner 5-fold cross-validation to maximize classification accuracy, usually yielding $C = 1$.

Security evaluation was then carried out on the test data. Similarly to the procedure used to tune λ , we manipulated each malicious sample according to Problem (8)-(9) for $c_{\max} \in [0, 20]$, assuming perfect (PK) and limited (LK) knowledge of the *true* discriminant function $g(\mathbf{x})$. As discussed in Sect. IV, in the LK case the attacker constructs evasion points by attacking a surrogate classifier $\hat{g}(\mathbf{x})$. These points are then used to attack the *true* classifier $g(\mathbf{x})$ and evaluate the performance. To provide a realistic evaluation, as in [19], we learn the surrogate classifier using a smaller training set, consisting of only 500 samples. For each value of c_{\max} we then computed TP at 1% FP for the true classifier under the PK and LK attack scenarios.

Experimental results. The average value (and standard deviation) of TP at 1% FP for the security evaluation procedure described above are reported in Fig. 2, for feature subset sizes of 100, 200, 300, and 400, and for the PK and LK attack scenarios. In the absence of attack (*i.e.*, when $c_{\max} = 0$), the two methods exhibited similar performances. Although the traditional method performed occasionally better than WAFS, the difference turned out not to be 95% statistically significant according to the Student’s t-test. Under attack (*i.e.*, when $c_{\max} > 0$), instead, WAFS always significantly outperformed the traditional method, for both the PK and the LK attack scenarios. As the performance of the traditional method decreased less gracefully as c_{\max} increased, we can conclude that WAFS leads to learning more secure classifiers. This is also confirmed by the LK attack scenario. In this case, even if only a surrogate classifier is available to the attacker, manipulating up to 20 words in a spam email may allow one to evade the true classifier almost surely (provided that the selected features are known to the attacker).

In Fig. 3, we also report the values of $G(\theta)$ (*i.e.*, the classification accuracy in the absence of attack) and $S(\theta)$ (*i.e.*, the average minimum number of features to be modified in a malicious sample to evade detection). Note how the proposed

²<http://spamassassin.apache.org/>, <http://spambayes.sourceforge.net/>

³Note that a similar procedure has also been carried out in [11], where it is also quantitatively shown that classification performance is not affected.

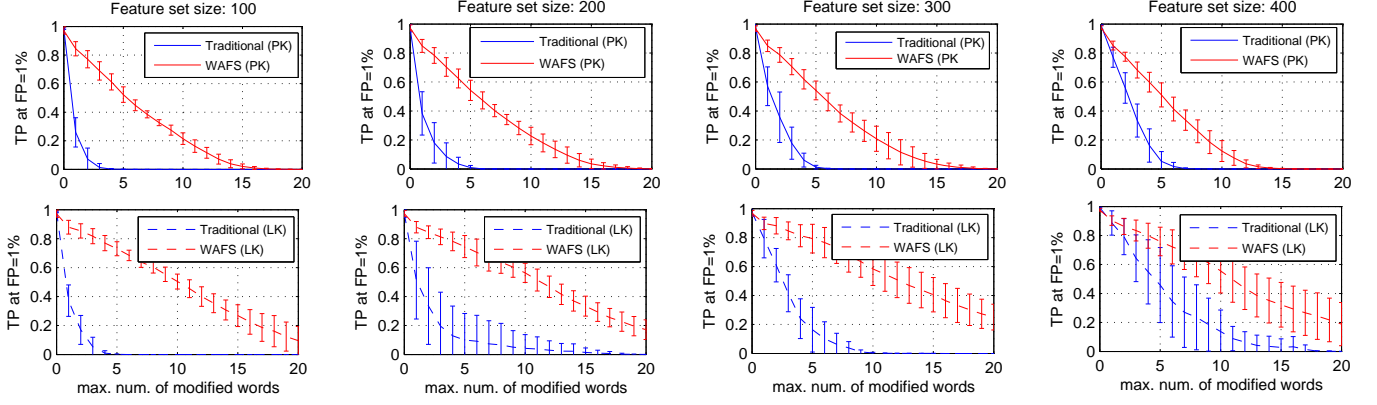


Fig. 2. Security evaluation curves for the spam data, showing the average and standard deviation of the TP rate at 1% FP rate, for feature subset sizes of 100, 200, 300, and 400 (in different columns), and for the PK (top row) and LK (bottom row) attack scenarios.

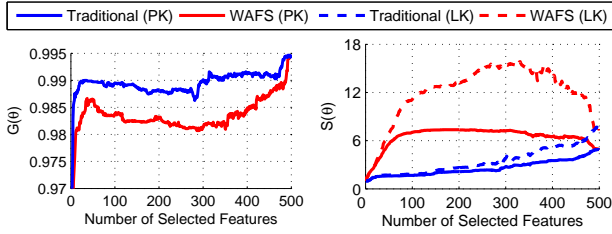


Fig. 3. Average classification accuracy $G(\theta)$ in the absence of attack (left plot), and classifier security $S(\theta)$ under attack (*i.e.*, average minimum number of modified words in each spam to evade detection) for the PK and LK attack scenarios (right plot), against different feature subset sizes, for the spam data.

adversarial feature selection method systematically required the attacker to modify a higher number of features (*i.e.*, words) to evade detection, for all considered feature subset sizes, without significantly affecting the accuracy in the absence of attack $G(\theta)$. This clearly confirms our previous results, showing that maximizing the security term $S(\theta)$ during feature selection helps improving the detection rate under attack, and thus classifier security. Although this comes at the cost of an increased computational complexity (since it requires simulating attacks against the targeted classifier), feature selection is often carried out offline, and thus the additional running time required by our method may not be critical.

Finally, we evaluated the correlation of the *hardness of evasion*, *i.e.*, the classifier security measure $S(\theta)$ used in this work, with the *weight evenness* [15], [16], *i.e.*, another possible measure for assessing the security of linear classifiers (see Sect. II-A). The goal is to verify whether the weight evenness can be adopted to compute the security term $S(\theta)$ in our approach, as it can be computed more efficiently than the hardness of evasion, *i.e.*, without simulating any attack against the trained classifier. To this end, we trained 200 linear classifiers using 200 distinct samples each, and evaluated the correlation between the two considered measures. Surprisingly, our experiment showed that the two measures were not significantly correlated (the Pearson's correlation coefficient was almost zero), contradicting the intuition in previous studies [15], [16]. One possible reason is that the weight evenness does not exploit any information on the data distribution besides the classifier's feature weights. Therefore, it may not be properly suited to characterize classifier security.

B. Malware Detection in PDF Files

Here, we consider another realistic application example related to the detection of malware (*i.e.*, *malicious software*) in PDF files. These files are characterized by a hierarchy of interconnected objects, each consisting of a *keyword*, denoting its type, and a *data stream*, representing its content; *e.g.*, the keyword `/PageLayout` characterizes an object describing how the corresponding page is formatted. This flexible, high-level structure allows for embedding of different kinds of content, such as JavaScript, Flash, and even binary programs, which in turn makes PDF files particularly attractive as vectors for disseminating malware. Recent work has exploited machine learning as a tool for detecting malicious PDF files based on their logical structure; in particular, on the set of embedded keywords [17], [18]. In this application example, we use the same feature representation exploited in [17], [19], in which each feature represents the number of occurrences of a given keyword in a PDF file. Conversely to the spam filtering example, feature values in this case can not be modified in an unconstrained manner to perform an evasion attack. In fact, it is not trivially possible to remove an embedded object (and the associated keywords) from a PDF without corrupting its structure. Nevertheless, it is quite easy to add new objects (*i.e.*, keywords) through the PDF versioning mechanism (see [17], [19] and references therein). In our case, this can be easily accounted for by setting $\mathbf{x} \leq \mathbf{x}'$ as an additional constraint to Problem (1)-(2) and Problem (8)-(9), respectively for the purpose of finding the optimal evasion points, and running the security evaluation procedure.⁴

Experimental setup. In these experiments, we considered the PDF malware dataset used in [17], [19], including 5591 legitimate and 5993 malicious PDFs. As mentioned above, features have integer values, each representing the occurrence of a given keyword in a PDF. In total, 114 distinct keywords were found from the first 1,000 samples (in chronological order). They were used as our set of features. As in [19], we limited the influence of outlying observations by setting the maximum value of each feature to 100. We then normalized each feature by simply dividing its value by 100. The SVM with the RBF kernel was used as the classification algorithm.

⁴With the notation $\mathbf{x} \leq \mathbf{x}'$, we mean that $x_j \leq x'_j$, for $j = 1, \dots, d$.

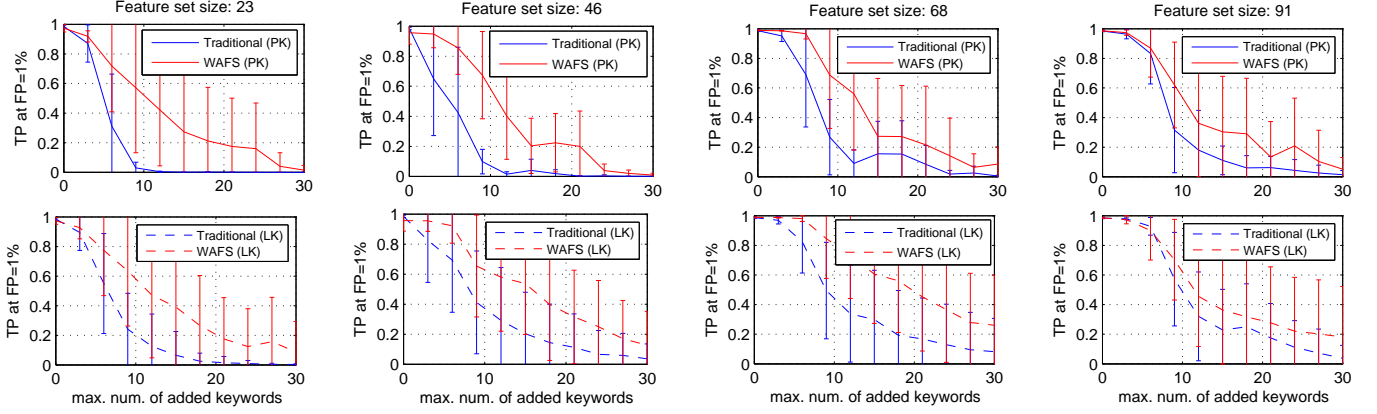


Fig. 4. Security evaluation curves for the PDF malware data, showing the average and standard deviation of the TP rate at 1% FP rate, for feature subset sizes of 23, 46, 68, and 91 (in different columns), and for the PK (top row) and LK (bottom row) attack scenarios.

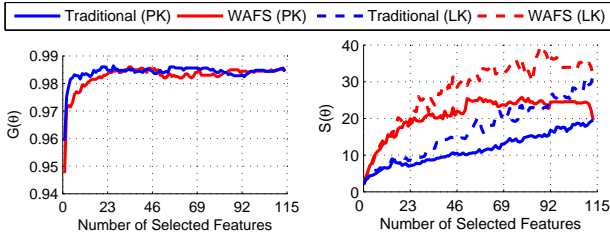


Fig. 5. Average classification accuracy $G(\theta)$ in the absence of attack (left plot), and classifier security $S(\theta)$ under attack (*i.e.*, average minimum number of added keywords to each malicious PDF to evade detection) for PK and LK attacks (right plot), against different feature subset sizes, for the PDF data.

The same performance $G(\theta)$ and classifier security $S(\theta)$ measures defined in the previous section were considered here. Notably, using the ℓ_1 -norm to evaluate $S(\theta)$ in this case amounts to counting the number of added keywords to a PDF (divided by 100). It is clear that the feature space is discrete in this case as well, since the admissible values are $\mathbf{x} \in \{0, 1/100, 2/100, \dots, 1\}^d$. We therefore exploited again Algorithm 2 on discrete spaces (Sect. III-B2) for the sake of estimating $S(\theta)$ and running security evaluation.

Following the same experimental setup used for the spam filtering task, we set $\lambda = 0.9$. Experiments were repeated ten times, each time randomly drawing 1,000 samples from the remaining data. In each run, these samples were split into a training and a test set of equal sizes. Then, feature subsets of sizes from 1 to 113 were selected using traditional and adversarial backward feature elimination, performing a 5-fold cross-validation on the training data. The SVM regularization parameter $C \in \{2^{-10}, 2^{-9}, \dots, 2^{10}\}$ and the kernel parameter $\gamma \in \{2^{-3}, 2^{-2}, \dots, 2^3\}$, were set during this process through an inner 5-fold cross-validation, maximizing the classification accuracy. This typically yielded $C = 256$ and $\gamma = 0.5$. Security evaluation was performed as for the previous experiments on spam filtering, considering $c_{\max} \in [0, 0.5]$, which amounts to adding a maximum of 50 keywords to each malicious PDF.

Experimental results. In Fig. 4, we report the average value (and standard deviation) of TP at 1% FP for the SVM with the RBF kernel trained on feature subsets of 23, 46, 68 and 91 features (*i.e.*, 20%, 40%, 60% and 80% of the total number of features), selected by the traditional backward elimination

wrapping algorithm and by the corresponding WAFS implementation, under the PK and LK attack scenarios. Similar observations to those reported for the spam filtering task can be made here; in particular, WAFS outperformed traditional feature selection in terms of TP values under attack, without exhibiting a significant performance loss in the absence of attack. Moreover, in the LK case, a higher number of added keywords was required, as expected, to reach a comparable detection rate to that reported for the PK attack scenario.

It is worth noting here that the variance of the TP rates reported in Fig. 4 turned out to be significantly higher than in the previous experiments. This fluctuation might be due to the use of smaller training sets, and to the higher variability induced by the use of a nonlinear decision boundary. Consequently, only few cases were 95% statistically significant based on the Student's t-test in the PK scenario. In the LK scenario, as the difference between the average values of the two methods was larger, WAFS was 95% significantly better than the traditional backward elimination algorithm in all cases for feature subsets of 23, 46 and 68 features, except for $c_{\max} \in [0, 0.03]$. Although some results were not 95% statistically significant due to the high variability of our results, we can nevertheless conclude that WAFS was able to outperform the traditional backward elimination approach in most of the cases.

The classification accuracy in the absence of attack $G(\theta)$, and the average minimum number of keywords added to a malicious PDF to evade detection (*i.e.*, $S(\theta) \times 100$) for the SVM with the RBF kernel trained on features selected by the traditional and the adversarial feature selection methods are shown in Fig. 5. Similarly to the results reported in the spam filtering example, the $S(\theta)$ values for WAFS are significantly higher than those exhibited by the traditional feature selection method in both the PK and the LK scenarios, although the SVM's classification accuracy $G(\theta)$ is not significantly affected. It should however be noted that the additional computational complexity required to compute $S(\theta)$ for a nonlinear classifier is higher than that required by a linear classifier due to the intrinsic complexity of exploring a nonlinear decision boundary. Finally, it is worth pointing out that WAFS was able to improve classifier security in this case by mainly selecting features that exhibited, on average,

higher values for the malicious class. In fact, due to the constraint $\mathbf{x} \leq \mathbf{x}'$, it becomes harder for an attacker to mimic characteristic feature values of the legitimate class in this case, yielding eventually a lower probability of evading detection. This may be an interesting research direction to explore, in order to devise surrogate measures of classifier security suitable for implementing adversarial feature selection as a more computationally efficient filter method.

VI. CONCLUSIONS AND FUTURE WORK

Feature selection may be considered a crucial step in security-related applications, such as spam and malware detection, when small subsets of features have to be selected to reduce computational complexity, or to improve classification performance by tackling the curse of dimensionality [38], [44]. However, since traditional feature selection methods implicitly assume that training and test samples follow the same underlying data distribution, their performance may be significantly affected under adversarial attacks that violate this assumption. Even worse, performing feature selection in adversarial settings may allow an attacker to evade the classifier at test time with a lower number of modifications to the malicious samples [11], [16], [35], [36]. To our knowledge, besides the above studies, the issue of selecting feature sets suitable for adversarial settings has neither been experimentally nor theoretically investigated more in depth.

In this paper, we proposed an *adversarial* feature selection method that optimizes not only the generalization capability of the wrapped classifier, but also its security against evasion attacks at test time. To this end, we extended a previous definition of classifier security, which was suited to linear classifiers trained on binary features, to the case of nonlinear classification algorithms trained on either continuous or discrete feature spaces. We validated the soundness of our approach on realistic application examples involving spam and PDF malware detection. We showed that the proposed approach can outperform traditional approaches in terms of classifier security, without significantly affecting the classifier performance in the absence of attacks. We also empirically showed that the proposed measure of classifier security provides a better characterization of this aspect than other previously-proposed measures aimed at evaluating the security of linear classifiers. However, our method demands for an increased computational complexity, with respect to traditional wrapping algorithms, as it requires simulating evasion attacks against the wrapped classifier at each iteration of the feature selection process. Although this may not be a critical aspect, as feature selection is often carried out offline, making our approach more efficient remains an open issue to be investigated in future work.

A possible solution to overcome this limitation, and exploit our method in the context of more efficient feature selection approaches like *filter* methods, may be to devise suitable surrogate measures of classifier security that can reliably approximate this value without simulating attacks against the trained classifier. Investigating the connections between security and stability of the feature selection process may be one fruitful research direction to this end, as discussed in Sect. II. A more concrete example is however given at the end

of Sect. V-B, based on the intuition of restricting the feasible space of sample manipulations available to the attacker. In practice, if the feature values of malicious samples can only be incremented, an adversarial feature selection procedure should prefer selecting features that exhibit lower values for samples in the legitimate class, making thus harder for an attacker to evade detection by mimicking such samples. This can be easily encoded by a measure that does not require training and attacking the corresponding classifier, and that can be thus exploited in the context of filter-based feature selection.

Another interesting extension of this work is related to the application of the proposed approach in the context of more complex feature mappings, *i.e.*, feature spaces in which there is not a clear, direct relationship with the characteristics of each sample, and therefore it is not trivial to understand how to modify a malicious sample to find an optimal evasion point, *i.e.*, to exhibit the desired feature values. This is however an application-specific issue, known in the adversarial machine learning literature as the inverse feature-mapping problem [10], [11]. In practice, the problem arises from the fact that optimal attacks are defined in feature space, and thus finding the corresponding optimal *real* samples requires reverse engineering the feature mapping. From a pragmatic perspective, this can be overcome by first defining a suitable set of manipulations that can be made to the real malicious samples (*e.g.*, many tools are available to obfuscate the content of malware samples, by manipulating their code [17]), and considering such manipulations as the only ones available to the attacker to find evasion points. Although this may lead us to find only suboptimal evasion points in feature space (as we restrict the attacker to work on a potentially smaller feasible set), we are guaranteed that the corresponding *real* samples not only exist, but they are also easily determined.

To conclude, we believe that our work provides a first, concrete attempt towards understanding the potential vulnerabilities of feature selection methods in adversarial settings, and towards developing more secure feature selection schemes against adversarial attacks.

ACKNOWLEDGMENTS

The authors would like to thank Davide Maiorca for providing them the PDF malware dataset. This work was supported in part by the National Natural Science Foundation of China under Grant 61003171, Grant 61272201, and Grant 61003172, and in part by Regione Autonoma della Sardegna under Grant CRP-18293, L. R. 7/2007, Bando 2009.

REFERENCES

- [1] B. Biggio, Z. Akhtar, G. Fumera, G. L. Marcialis, and F. Roli, "Security evaluation of biometric authentication systems under real spoofing attacks," *IET Biometrics*, vol. 1, no. 1, pp. 11–24, 2012.
- [2] R. N. Rodrigues, L. L. Ling, and V. Govindaraju, "Robustness of multimodal biometric fusion methods against spoof attacks," *J. Vis. Lang. Comput.*, vol. 20, no. 3, pp. 169–179, 2009.
- [3] J. Newsome, B. Karp, and D. Song, "Paragraph: Thwarting signature learning by training maliciously," in *RAID*, ser. LNCS. Springer, 2006, pp. 81–105.
- [4] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage, "Inferring internet denial-of-service activity," *ACM Trans. Comput. Syst.*, vol. 24, no. 2, pp. 115–139, 2006.

- [5] D. Lowd and C. Meek, "Good word attacks on statistical spam filters," in *2nd Conf. on Email and Anti-Spam*, CA, USA, 2005, pp. 87–94.
- [6] B. Nelson, M. Barreno, F. Jack Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar, and K. Xia, "Misleading learners: Co-opting your spam filter," in *Mach. Learn. in Cyber Trust*. Springer US, 2009, pp. 17–51.
- [7] B. Biggio, I. Corona, B. Nelson, B. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli, "Security evaluation of support vector machines in adversarial environments," in *Supp. Vect. Mach. Apps.*. Springer Int'l Publishing, 2014, pp. 105–153.
- [8] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *10th ACM SIGKDD Int'l Conf. on Knowl. Discovery and Data Mining*, 2004, pp. 99–108.
- [9] B. Biggio, G. Fumera, and F. Roli, "Design of robust classifiers for adversarial environments," in *IEEE Int'l Conf. on Syst., Man, and Cyb.*, 2011, pp. 977–982.
- [10] L. Huang, A. D. Joseph, B. Nelson, B. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *4th ACM Workshop on Artificial Intell. and Security*, 2011, pp. 43–57.
- [11] B. Biggio, G. Fumera, and F. Roli, "Security evaluation of pattern classifiers under attack," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 4, pp. 984–996, 2014.
- [12] M. Brückner, C. Kanzow, and T. Scheffer, "Static prediction games for adversarial learning problems," *JMLR*, vol. 13, pp. 2617–2654, 2012.
- [13] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *ASIACCS '06: Proc. 2006 ACM Symp. on Info., Comput. and Comm. Sec.*. ACM, 2006, pp. 16–25.
- [14] G. L. Wittel and S. F. Wu, "On attacking statistical spam filters," in *1st Conf. on Email and Anti-Spam*, CA, USA, 2004.
- [15] A. Kolcz and C. H. Teo, "Feature weighting for improved classifier robustness," in *6th Conf. on Email and Anti-Spam*, CA, USA, 2009.
- [16] B. Biggio, G. Fumera, and F. Roli, "Multiple classifier systems for robust classifier design in adversarial environments," *Int'l J. Mach. Learn. Cyb.*, vol. 1, no. 1, pp. 27–41, 2010.
- [17] D. Maiorca, I. Corona, and G. Giacinto, "Looking at the bag is not enough to find the bomb: an evasion of structural methods for malicious pdf files detection," in *Proc. 8th ACM SIGSAC Symp. on Info., Comp. and Comm. Sec.*, ser. ASIACCS '13. ACM, 2013, pp. 119–130.
- [18] N. Šrđić and P. Laskov, "Detection of malicious pdf files based on hierarchical document structure," in *Proc. 20th Annual Network & Distributed Syst. Sec. Symp.*. The Internet Society, 2013.
- [19] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrđić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *ECML PKDD, Part III*, ser. LNCS, vol. 8190. Springer Berlin Heidelberg, 2013, pp. 387–402.
- [20] A. Globerson and S. T. Roweis, "Nightmare at test time: robust learning by feature deletion," in *Proc. 23rd Int'l Conf. on Mach. Learn.*, W. W. Cohen and A. Moore, Eds., vol. 148. ACM, 2006, pp. 353–360.
- [21] B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, and F. Roli, "Is data clustering in adversarial settings secure?" in *AISeC'13: Proc. Artificial Intell. and Sec. Workshop*. ACM, 2013, pp. 87–98.
- [22] B. Biggio, S. R. Bulò, I. Pillai, M. Mura, E. Z. Mequanint, M. Pelillo, and F. Roli, "Poisoning complete-linkage hierarchical clustering," in *Structural, Syntactic, and Statistical Patt. Rec.*, ser. LNCS, vol. 8621. Springer, 2014, pp. 42–52.
- [23] Z. Jorgensen, Y. Zhou, and M. Inge, "A multiple instance learning strategy for combating good word attacks on spam filters," *JMLR*, vol. 9, pp. 1115–1146, 2008.
- [24] H. Lee and A. Y. Ng, "Spam deobfuscation using a hidden Markov model," in *Int'l Conf. Email and Anti-Spam*, 2005.
- [25] K. Wang, J. J. Parekh, and S. J. Stolfo, "Anagram: A content anomaly detector resistant to mimicry attack," in *RAID'06*, 2006, pp. 226–248.
- [26] B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, and J. D. Tygar, "Antidote: understanding and defending against poisoning of anomaly detectors," in *9th Internet Meas. Conf.*, ser. IMC '09. ACM, 2009, pp. 1–14.
- [27] M. Bishop, J. Cummins, S. Peisert, A. Singh, B. Bhuniratan, D. Agarwal, D. Frincke, and M. Hogarth, "Relationships and data sanitization: A study in scarlet," in *W. New Sec. Paradigms*. ACM, 2010, pp. 151–164.
- [28] M. Kloft and P. Laskov, "Online anomaly detection under adversarial impact," in *13th Int'l Conf. on AI and Statistics*, 2010, pp. 405–412.
- [29] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *29th Int'l Conf. Mach. Learn.*. Omnipress, 2012, pp. 1807–1814.
- [30] C. H. Teo, A. Globerson, S. Roweis, and A. Smola, "Convex learning with invariances," in *NIPS 20*. Cambridge, MA: MIT Press, 2008, pp. 1489–1496.
- [31] B. Biggio, G. Fumera, and F. Roli, "Multiple classifier systems under attack," in *MCS*, ser. LNCS, vol. 5997. Springer, 2010, pp. 74–83.
- [32] G. F. Cretu, A. Stavrou, M. E. Locasto, S. J. Stolfo, and A. D. Keromytis, "Casting out demons: Sanitizing training data for anomaly sensors," in *IEEE Symp. on Sec. and Privacy*. IEEE CS, 2008, pp. 81–95.
- [33] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, "Bagging classifiers for fighting poisoning attacks in adversarial environments," in *MCS*, ser. LNCS, vol. 6713. Springer, 2011, pp. 350–359.
- [34] B. Biggio, G. Fumera, and F. Roli, "Evade hard multiple classifier systems," in *Supervised and Unsupervised Ensemble Methods and Their Applications*, ser. Studies in Comp. Intell., Springer Berlin / Heidelberg, 2009, vol. 245, pp. 15–38.
- [35] B. Li and Y. Vorobeychik, "Feature cross-substitution in adversarial classification," in *NIPS 27*. Curran Associates, Inc., 2014, pp. 2087–2095.
- [36] F. Wang, W. Liu, and S. Chawla, "On sparse feature attacks in adversarial learning," in *IEEE Int'l Conf. on Data Mining (ICDM)*. IEEE, 2014, pp. 1013–1018.
- [37] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, 2002.
- [38] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *JMLR*, vol. 13, pp. 27–66, 2012.
- [39] M. Barreno, B. Nelson, A. Joseph, and J. Tygar, "The security of machine learning," *Mach. Learn.*, vol. 81, pp. 121–148, 2010.
- [40] M. Kloft and P. Laskov, "A 'poisoning' attack against online anomaly detection," in *NIPS*, 2007.
- [41] D. Lowd and C. Meek, "Adversarial learning," in *KDD*, Chicago, IL., 2005, pp. 641–647.
- [42] B. Nelson, B. I. Rubinstein, L. Huang, A. D. Joseph, S. J. Lee, S. Rao, and J. D. Tygar, "Query strategies for evading convex-inducing classifiers," *JMLR*, vol. 13, pp. 1293–1332, 2012.
- [43] B. Biggio, G. Fumera, and F. Roli, "Multiple classifier systems for adversarial classification tasks," in *MCS*, ser. LNCS, vol. 5519. Springer, 2009, pp. 132–141.
- [44] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *JMLR*, vol. 3, pp. 1157–1182, 2003.
- [45] M. Kolar and H. Liu, "Feature selection in high-dimensional classification," *JMLR (ICML Track)*, vol. 28, pp. 329–337, 2013.
- [46] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [47] Q. Liu, A. H. Sung, M. Qiao, Z. Chen, and B. Ribeiro, "An improved approach to steganalysis of JPEG images," *Info. Sciences*, vol. 180, no. 9, pp. 1643 – 1655, 2010.
- [48] Q. Liu, A. H. Sung, Z. Chen, and J. Xu, "Feature mining and pattern classification for steganalysis of LSB matching steganography in grayscale images," *Pattern Recognition*, vol. 41, pp. 56–66, 2008.
- [49] W. Lee, W. Fan, M. Miller, S. J. Stolfo, and E. Zadok, "Toward cost-sensitive modeling for intrusion detection and response," *J. Comput. Sec.*, vol. 10, no. 1–2, pp. 5–22, 2002.
- [50] M.-J. Tsai, C.-S. Wang, J. Liu, and J.-S. Yin, "Using decision fusion of feature selection in digital forensics for camera source model identification," *Comp. Standards & Interf.*, vol. 34, no. 3, pp. 292–304, 2012.
- [51] S. M. Lee, D. S. Kim, J. H. Kim, and J. S. Park, "Spam detection using feature selection and parameters optimization," in *Int'l Conf. on Complex, Intell. and Software Intensive Syst.*, 2010, pp. 883–888.
- [52] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial Intell.*, vol. 151, no. 1, pp. 155–176, 2003.
- [53] D. Andrea and A. Nicholas, "Feature selection via probabilistic outputs," in *29th Int'l Conf. Mach. Learn.*, pp. 1791–1798, 2012.
- [54] P. Maji and P. Garai, "Fuzzy-rough simultaneous attribute selection and feature extraction algorithm," *IEEE Trans. Cyb.*, vol. 43, no. 4, pp. 1166–1177, 2013.
- [55] R. Diao and Q. Shen, "Feature selection with harmony search," *Syst., Man, and Cyb., Part B: IEEE Trans. Cyb.*, vol. 42, no. 6, pp. 1509–1523, 2012.
- [56] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [57] J.-B. Yang and C.-J. Ong, "An effective feature selection method via mutual information estimation," *Syst., Man, and Cyb., Part B: IEEE Trans. Cyb.*, vol. 42, no. 6, pp. 1550–1559, 2012.
- [58] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intell.*, vol. 97, no. 1, pp. 273–324, 1997.

- [59] J. Neumann, C. Schnörr, and G. Steidl, "Combined SVM-based feature selection and classification," *Mach. Learn.*, vol. 61, no. 1-3, pp. 129–150, 2005.
- [60] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *J. of Mach. Learn. Res.*, vol. 3, pp. 1439–1461, 2003.
- [61] H. A. Le Thi, X. T. Vo, and T. P. Dinh, "Robust feature selection for SVMs under uncertain data," in *Advances in Data Mining, Apps, and Theoretical Aspects*. Springer, 2013, pp. 151–165.
- [62] P. Huber, *Robust Statistics*, ser. Probability and Mathematical Statistics. New York, NY, USA: John Wiley and Sons, 1981.
- [63] R. A. Maronna, R. D. Martin, and V. J. Yohai, *Robust Statistics: Theory and Methods*, ser. Probability and Mathematical Statistics. New York, NY, USA: John Wiley and Sons, 2006.
- [64] C. Croux, P. Filzmoser, and M. R. Oliveira, "Algorithms for projection - pursuit robust principal component analysis," *Chemometrics and Intell. Lab. Syst.*, vol. 87, no. 2, pp. 218–225, 2007.
- [65] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.
- [66] C. D. Manning and H. Schütze, *Foundations of statistical natural lang. proc.*. MIT Press, 1999, vol. 999.
- [67] G. V. Cormack, "TREC 2007 spam track overview," in *TREC*, E. M. Voorhees and L. P. Buckland, Eds., vol. SP 500-274. NIST, 2007.
- [68] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.



Daniel S. Yeung (F'04) received his Ph.D. from Case Western Reserve University, USA. He was a faculty at Rochester Institute of Technology from 74-80. In the next ten years he held industrial and business positions in USA. In 89 he joined City Polytechnic of Hong Kong as an Associate Head/Principal Lecturer at the Department of Computer Science. Then he served as the founding Head and Chair Professor of the Department of Computing at The Hong Kong Polytechnic University until his retirement at 06. Currently he is a Visiting Professor in the School of Computer Science and Engineering, South China University of Technology. Dr. Yeung is a fellow of the IEEE and served as the President of IEEE SMC Society in 08-09. His current research interests include neural-network sensitivity analysis, large scale data retrieval problem and cyber security.



Fei Zhang (S'11) received her B.S. degree in Information and Computer Science from Minnan Normal University, Zhangzhou, China, in 2009. Now she is a Ph. D. student in South China University of Technology. Her current interested research is focused on machine learning, computer security and spam filtering. Miss Zhang is an IEEE student member.



Patrick P.K. Chan (M'04) received the Ph.D. degree from Hong Kong Polytechnic University in 09. He is currently Associate Professor of School of Computer Science and Engineering in South China University of Technology, China. His current research interests include pattern recognition, adversarial learning, and multiple classifier systems. Dr. Chan is a member of the governing boards of IEEE SMC Society 14-16. He is also the Chairman of IEEE SMCS Hong Kong Chapter 14-15 and the counselor of IEEE Student Branch in South China

University of Technology.



Fabio Roli (F'12) received his Ph.D. in Electronic Eng. from the Univ. of Genoa, Italy. He was a research group member of the Univ. of Genoa (88-94). He was adjunct professor at the University of Trento (93-94). In 95, he joined the Dept. of Electrical and Electronic Eng. of the Univ. of Cagliari, where he is now professor of Computer Eng. and head of the research laboratory on pattern recognition and applications. His research activity is focused on the design of pattern recognition systems and their applications. He was a very active organizer of int'l conferences and workshops, and established the popular workshop series on multiple classifier systems. Dr. Roli is Fellow of the IEEE and of the Int'l Association for Pattern Recognition.



Battista Biggio (M'07) received the M.Sc. degree (Hons.) in electronic engineering and the Ph.D. degree in electronic engineering and computer science from the University of Cagliari, Italy, in 2006 and 2010. Since 2007, he has been with the Department of Electrical and Electronic Engineering, University of Cagliari, where he is currently a post-doctoral researcher. In 2011, he visited the University of Tbingen, Germany, and worked on the security of machine learning to training data poisoning. His research interests include secure machine learning, multiple classifier systems, kernel methods, biometrics and computer security. Dr. Biggio serves as a reviewer for several international conferences and journals. He is a member of the IEEE and of the IAPR.