Randomizing SVM against Adversarial Attacks Under Uncertainty

Yan Chen[1⋆], Wei Wang[2] and Xiangliang Zhang[3(✉)]

[1] Columbia University, New York, USA
yc3107@columbia.edu
[2] Beijing Jiaotong University, Beijing, China
wei.wang.email@gmail.com
[3] King Abdullah University of Science and Technology (KAUST), Saudi Arabia
xiangliang.zhang@kaust.edu.sa

Abstract. *Robust machine learning algorithms have been widely studied in adversarial environments where the adversary maliciously manipulates data samples to evade security systems. In this paper, we propose randomized SVMs against generalized adversarial attacks under uncertainty, through learning a classifier distribution rather than a single classifier in traditional robust SVMs. The randomized SVMs have advantages on better resistance against attacks while preserving high accuracy of classification, especially for non-separable cases. The experimental results demonstrate the effectiveness of our proposed models on defending against various attacks, including aggressive attacks with uncertainty.*

Keywords: Adversarial Learning, Robust SVM, Randomization

1 Introduction

Adversary machine learning is an important research track that harnesses machine learning to resolve security issues. The adversary can deliberately manipulate their data to mislead the defender of security. Machine learning is challenged by learning from poisoned training data [2][7]. Consequently, it is imperative to identify potential vulnerabilities and propose countermeasures in order to improve the robustness of machine learning algorithms against attacks [9][6][8].

Support Vector Machines (SVMs) as supervised models are among the most popular classification techniques adopted in security applications like malware detection, intrusion detection, and spam filtering [13,4]. In order to secure decision-making system against *poisoning attacks* (contaminating training data), *Robust SVMs* as the modification to standard SVMs had been proposed by exploring robustness, kernels and dual formulations in SVMs and Bayes learning [12],[11], [13], [10]. In general, the intuition is to make the decision boundary learned in robust SVMs not be extremely sensitive to any single training example. Recently,

⋆ The work was completed when the first author was visiting KAUST as an intern.

Randomized SVMs are studied for defending a classifier against *exploratory attacks*, which probe the classifier with queries in order to reveal confidential information about the training dataset [1]. Randomized SVMs aim at learning *a distribution of classifiers*, rather than a single classifier in previous study of robust SVMs, and thus make the system less vulnerable.

In this paper, we study how to design the *randomized SVMs* against *poisoning attacks with uncertainty*, which are more sophisticated than previously studied attacks, e.g., free range and restrained range attack in [13]. The idea of randomized SVM is demonstrated in Fig. 1. Standard SVM linear classifier learns a w that separates positive and negative class with maximal margin (Fig. 1 (a)). Randomized SVM learns a distribution about w, for example, a Gaussian distribution $\mathcal{D}_w = \mathcal{N}(u, \Sigma)$, demonstrated in Fig. 1 (c). Such a distribution can guarantee the classification accuracy of w sampled from $\mathcal{D}_w$ with a separability higher than ν (the probability that training data can be separated), i.e., $\mathbb{P}_{w\sim\mathcal{N}(u,\Sigma)}(y_i(w^T x_i) \geq 1) \geq \nu$. In well-separated cases, classifier distribution is as good as a single classifier, but provides a set of classifiers with the same performance to confuse attackers when they attempt to understand the classification system and prepare attacks. In the case where the adversary adds noise to mislead the system, the region close to decision boundary usually becomes complicated. A deterministic classifier has to separate all data with probability 1, and thus accuracy is scarified. Randomized classifier lowers the separation standard (with probability ν less than 1) and guarantees the accuracy, as shown in Fig. 1 (d). Therefore, we investigate randomized SVM as a promising solution to learn robust classifiers against poisoning attacks.

The main contributions of our work are:

1) We design randomized SVM models against different types of attacks and formulate each model into convex optimization, second order cone programming (SOCP) or semi-definite programming (SDP);

2) The attacks we define to challenge randomized SVMs are generalized from previously studied restrained range (RR) attacks. The generalized attacks (RR with uncertainty, and distributional range attack with uncertainty) are more aggressive and complicated, and cover a wide range of attacks;

3) We evaluate our randomized SVMs on several data sets. The experimental results show that *randomized* SVMs outperform existing *robust* SVMs on defending against attacks at different level of intensity.

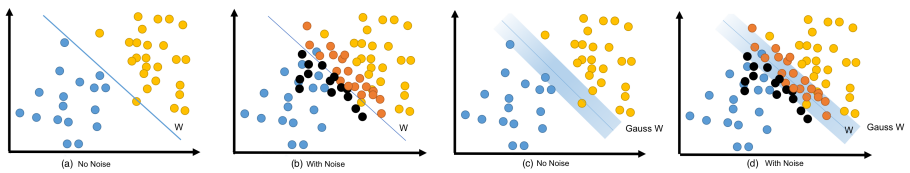


Fig. 1: Standard classifier (a,b) and randomized classifiers (c,d) when input without (a,c) and with noise (b,d) shown in black and orange.

2 Related Work

[4] comprehensively evaluated SVMs in adversarial environment, i.e., how SVMs cope with different types of attacks, such as poisoning (contaminating the training data), envision (circumventing the learned classifier) and privacy-breaching attacks. Our work focuses on learning optimal SVMs against poisoning attacks. Therefore, we first discuss the related work that also studies the learning of SVMs from poisoned training data, where each sample is manipulated by adding a noise Δx, i.e., $x' = x + \Delta x$.

There are many Robust SVM models modified from the standard SVM for handling noise in training data. [3] formulate SVM learning with contaminated training data by modeling the unobserved true input (uncorrupted data) as a hidden mixture component. The added noise is assumed to be bounded as $||\Delta x_i|| \leq \delta_i$, such that a noisy sample lies within a ball of given radius w.r.t. the true non-noisy sample. [11] prove that such SVMs with norm-based regularization build in a robustness to sample noise whose probability level sets are symmetric unit balls with respect to the dual of the regularizing norm. Different but relevant work in [12],[5] studies how SVMs are affected by adversarial label noise (e.g., flipping labels of certain training samples), rather than by feature noise (e.g., adding noise to training samples).

The most relevant work in [13] develops robust SVMs against two attack models: free range and restrained range, which are more realistic manipulations attackers can make, while the noise bound δ_i in [3] is fixed and known to both attackers and defenders. In this paper, we study generalized attack models of restrained range (RR), which is more advantageous than free range attacks. The generalized attacks have more flexibility on designing attacks in different forms and with more uncertainty.

The other stream of related work is Randomized SVMs. [1] investigate randomization as a suitable strategy for protecting SVMs against exploratory attacks. Unlike poisoning attacks, exploratory attacks occur after the training stage and aim at revealing classification boundary by probing with queries. To protect the classification system, instead of learning a fixed classifier, the defender uses training data to infer a distribution of classifiers. The decision system is thus not deterministic but probabilistic. In this paper, we develop randomized SVMs for RR attacks, and also two generalized attack models with more uncertainty.

3 Problem Definition

Denote a training sample x_i where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$ as the label of x_i, $i = 1, ..., n$. We consider the adversary learning problem where the adversary aims to modify training data in the feature space to their desired targets to mislead the classifier learning. For example in spam detection, the x_i includes some demonstrative information from individual email while y_i is the indicator that judges if the email is malicious. For a malicious sample x_i with $y_i = 1$, the adversary can modify it to be $x_i + \delta_i$ for avoiding detection and misleading the

classifier training process according to his planned goals. For the same example in spam detection, good words can be added to a spam email to defeat spam detectors. Following the assumption in [13], the adversary does not modify the innocuous data (with $y_i = -1$), e.g., the adversary has no intention to modify legitimate e-mails.

Our target is to learn a Gaussian distribution $\mathcal{D}_w = \mathcal{N}(u, \Sigma)$, where each sample w is a classifier for discriminating x_i with different labels. Given a required separable probability ν ($\mathbb{P}_{w\sim\mathcal{N}(u,\Sigma)}(y_i(w^T x_i) \geq 1) \geq \nu$), such a distribution can guarantee the classification accuracy of w sampled from $\mathcal{D}_w$. In simple words, it correctly differentiates positive and negative samples as many as possible, while allowing some samples be separable with a low probability. When training data are poisoned, our randomized SVMs are able to classify innocuous data correctly while allowing samples close to the decision boundary (that are probably manipulated samples with noise) be separable with a low probability. Therefore, randomized SVMs are expected to be more robust against poisoning attacks than deterministic SVMs.

4 Attack Model Design

[13] introduces a Restrained Range (RR) Attack model, which allows modification of x_i in a limited range, as a large modification of original x_i entails loss of malicious utility. The modification of x_i is proportional to the difference between x_i and x_i^t (the target of modifying x_i), and is usually set according to the adversary's estimate of the innocuous data. We generalize RR attack with uncertainty in different norm settings. The adversary will not only have the freedom to move data in the feature space, but also can develop attacks with different range shape. Then a most general attack model is defined by considering that the adversary probably manipulates deliberately data with uncertain distribution, unknown expectation and variance to develop an infinite dimensional attack space.

4.1 Restrained Range Attack with Uncertainty

The restrained attack in [13] is defined as

$$0 \leq (x_{ij}^t - x_{ij})\delta_{ij} \leq c_f(1 - \frac{|x_{ij}^t - x_{ij}|}{|x_{ij}| + |x_{ij}^t|})(x_{ij}^t - x_{ij})^2. \tag{1}$$

Dividing by $(x_{ij}^t - x_{ij})$, we obtain the bound of δ_{ij}

$$0 \leq |\delta_{ij}| \leq c_f(1 - \frac{|x_{ij}^t - x_{ij}|}{|x_{ij}| + |x_{ij}^t|})(|x_{ij}^t - x_{ij}|). \tag{2}$$

We generalize the above bound for δ_i as

$$\delta_i = \{\delta_i : P_i r, ||r|| \leq c_f||v_i||, v_{ij} = x_{ij}^t - x_{ij}\} \tag{3}$$

which provides more freedom to set the shrink matrix P. When setting P as $(1 - \frac{|x_{ij}^t - x_{ij}|}{|x_{ij}| + |x_{ij}^t|})$ and implementing L_1 norm, the generalized restrained attack is

approximately reduced to prior form. The main difficult in restrained attacks is to estimate the target x_{ij}^t. The defender usually utilizes their prior knowledge to guess the most possible x_{ij}^t. In fact, the adversary is reluctant to make target data move far away from the origin, which leading to loss of maliciousness. A simple method to estimate x_{ij}^t is to calculate the mean and variance of malicious data to obtain $x_{ij}^t = x_{ij} + c_\delta \varepsilon_{ij}$, where c_δ is the standard variation of these samples and ε_{ij} is a random noise.

4.2 Distributional Range Attack with Uncertainty

To introduce an infinite-dimensional uncertain set to attack models, we define a most general attack model where the modification δ_i follows a distribution belonging to the set:

$$\Delta = \{p : supp(p) = R^n, E_p[\delta_i] = m_i, E_p[||\delta_i||] \leq \sigma_i\}. \tag{4}$$

This attack model processes remarkably probability uncertainty. The parameter m is the central point that attacks may happen. According to strong law of large number, we know when the aggressive samples are sufficient large, the average results are close to the expectation of attacks. Thus, m and σ control the intensity of attacks, similar to c_f in (3). The above attack model is the most generalized result by considering all variables in the probability measure space. When the mean is properly set and the variance is sufficient large, it can be implemented to cover RR models. It is expected that by solving such a problem we would obtain an optimal classifier against all distributional attacks.

5 Randomized SVMs Learning

For the different attack strategies defined above, we develop randomized classifiers for learning against noise. Randomized SVMs introduced in [1] learn a distribution $\mathcal{D}_w = \mathcal{N}(u, \Sigma)$, from which a randomly drawn w can make a system of linear inequalities $y_i(w^T x_i) \geq 1$ satisfied with a probability that exceeds ν, where $0 \leq \nu \leq 1$. That is to say, the probability of data samples that are separable by D_w is at least ν,

$$\mathbb{P}_{w \sim \mathcal{N}(u,\Sigma)}(y_i(w^T x_i) \geq 1) \geq \nu. \tag{5}$$

The optimization problem of identifying the parameters of $\mathcal{D}_w$ (u and $\Sigma = s * s^T$) is defined as:

$$\begin{aligned} &\min_{u,\xi,s} \ \frac{1}{2}\frac{u^T u}{1^T s} + C\sum_{i=1}^{n}\xi_i \\ &s.t. \ y_i^T(u^T x_i) \geq 1 + \Phi^{-1}(\nu)\sum_{j=1}^{d} x_{ij}^2 s_j - \xi_i, \\ &\qquad s_i \geq 0, i = 1, 2, ..., d. \qquad \xi_i \geq 0, i = 1, 2, ..., n. \end{aligned} \tag{6}$$

We now formulate randomized SVM for the above-mentioned different types of attacks, including the Restrained Range Attack (RRA), Restrained Range Attack with Uncertainty (RRAU), and Distributional Range Attack with Uncertainty (DRAU). They all modify x_i with $y_i = 1$ to be $x_i+\delta_i$, without changing x_i with $y_i = -1$. The hinge loss of classification can be defined as

$$h(w, b, x_i) = \begin{cases} \max_{\delta_i} \max(1 - (w^T(x_i + \delta_i) + b), 0) & \text{for } y_i = 1; \\ \max(1 + w^T x_i + b, 0) & \text{for } y_i = -1. \end{cases}$$

Combing these two loss functions, the objective function in (6) becomes

$$\min_{w,b} \sum_i \max_{\delta_i} (1 - y_i(w^T x_i + b) - \frac{1}{2}(1 + y_i)w^T \delta_i)^+ + \frac{1}{2}\frac{u^T u}{1^T s}. \tag{7}$$

and can be simplified as

$$\min_{w,b} \sum_i \max_{\delta_i} (1 - y_i(w^T x_i) - \frac{1}{2}(1 + y_i)u^T \delta_i)^+ + \frac{1}{2}\frac{u^T u}{1^T s}. \tag{8}$$

The unique term in (8) relevant with δ_i is $\max_{\delta_i}(-\frac{1}{2}(1 + y_i)u^T\delta_i)$ for fixed u. Then for the following model derivation, we will focus on the sub-problem

$$\max_{\delta_i}(-\frac{1}{2}(1 + y_i)u^T\delta_i) \tag{9}$$

with different constraints in different attack models.

5.1 Randomized SVM against RRA

RRA sets δ_i by (1). Let $e_{ij} = c_f(1 - \frac{|x_{ij}^t - x_{ij}|}{|x_{ij}+|x_{ij}^t|})(x_{ij}^t - x_{ij})^2$. The sub-problem relevant with δ_i in (9) becomes

$$\begin{aligned} &\max_{\delta_i} -\frac{1}{2}(1 + y_i)u^T\delta_i \\ &s.t.\ 0 \le (x_{ij}^t - x_{ij})\delta_{ij} \le e_{ij}. \end{aligned} \tag{10}$$

Introducing $(-z_{ij} + v_{ij})(x_{ij}^t - x_{ij}) = \frac{1}{2}(1 + y_i)u_j, z_i \succeq 0, v_i \succeq 0$ into the above problem, it is transformed into $\max -\sum_j(-z_{ij}e_{ij} + v_{ij}0)$, by solving which, we can rewrite the optimization problem in (6) for Randomized SVM against RRA as follows,

$$\begin{aligned} &\min\ \max_{u,s} \frac{u^T u}{1^T s} + C\sum_i \xi_i \\ &s.t.\ \xi_i \ge 0, z_i \succeq 0, v_i \succeq 0, t_i \ge \sum_j (z_{ij}e_{ij}), i = 1, ..., n,\ s_i \ge 0, i = 1, 2, ..., d, \\ &y_i^T(u^T x_i) \ge 1 + \Phi^{-1}(\nu)\sum_{j=1}^{d} x_{ij}^2 s_j - \xi_i + t_i,\ \ (-z_{ij} + v_{ij})(x_{ij}^t - x_{ij}) = \frac{1}{2}(1 + y_i)u_j. \end{aligned} \tag{11}$$

All constraints in the formulation are linear, we would obtain similar SDP by considering the worst situation.

5.2 Randomized SVM against RRAU

RRAU sets δ_i by (3). Similarly, we have the sub-problem

$$\begin{aligned} &\max_{\delta_i} -\frac{1}{2}(1+y_i)u^T\delta_i \\ &s.t. \quad \delta_i = \{\delta_i : P_i r, ||r|| \le c_f||v_i||, v_{ij} = x_{ij}^t - x_{ij}\}. \end{aligned} \tag{12}$$

Let $f_i = \max -\frac{1}{2}(1+y_i)u^T P_i r,\ s.t.\ ||r|| \le c_f||v_i||$. Since we have $f_i \le \frac{1}{2}(1+y_i)||v_i||||P_i u||_*$. If the norm of v is L_1 norm, the optimization problem for Randomized SVM against RRAU as follows,

$$\begin{aligned} &\min\ \max_{u,s} \frac{u^T u}{1^T s} + C\sum_i \xi_i \\ &s.t.\ \xi_i \ge 0, i = 1,2,...,n, \qquad s_i \ge 0, i = 1,2,...,d, \\ &\qquad y_i^T(u^T x_i) \ge 1 + \Phi^{-1}(\nu)\sum_{j=1}^{d} x_{ij}^2 s_j - \xi_i + t_i, \\ &t_i \ge \frac{1}{2}(1+y_i)c_f\sum_{j=1}^{d}|x_{ij}^t - x_{ij}|||P_i u||_*, i = 1,2,...,n. \end{aligned} \tag{13}$$

5.3 Randomized SVM against DRAU

DRAU sets δ_i by (4). The original optimization problem described in (8) with DRAU becomes

$$\min_{w,b}\sum_i \max_{\delta_i \in \Delta} E_p(1 - y_i(w^T x_i + b) - \frac{1}{2}(1+y_i)w^T\delta_i)^+ + \frac{1}{2}\frac{u^T u}{1^T s}. \tag{14}$$

We first consider the maximizing inner optimization problem,

$$\begin{aligned} &\max_p E_p(1 - y_i(w^T x_i + b) - \frac{1}{2}(1+y_i)w^T\delta_i)^+ \\ &s.t.\ E_p(\delta_i) = m_i,\ E_p(||\delta_i||) \le \sigma_i,\ i = 1,2,...,n. \end{aligned} \tag{15}$$

The *dual Lagrange function* of it can be written as $g(\lambda_1,\lambda_2) = \lambda_1 m_i + \lambda_2\sigma_i + \max_p\{E_p(1 - y_i(w^T x_i + b) - \frac{1}{2}(1+y_i)w^T\delta_i)^+ - \lambda_1\delta_i - \lambda_2||\delta_i||\}$, which can be equivalently written as $g(\lambda_1,\lambda_2) = \lambda_1 m_i + \lambda_2\sigma_i + \max_{z_i}\{(1 - y_i(w^T x_i + b) - \frac{1}{2}(1+y_i)w^T z_i)^+ - \lambda_1 z_i - \lambda_2||z_i||\}$. Since $x^+ = max(x,0)$, it is further written as

$$\begin{aligned} g(\lambda_1,\lambda_2) = \lambda_1 m_i + \lambda_2\sigma_i + \max\{&\max_{z_i}[(1 - y_i(w^T x_i + b) \\ &-\frac{1}{2}(1+y_i)(w+\lambda_1)^T z_i)^+ - \lambda_2||z_i||], \max_{z_i}[-\lambda_1^T z_i - \lambda_2||z_i||]\}. \end{aligned}$$

By Cauchy-Schwarz inequality, note $(w+\lambda_1)^T z_i \le ||w+\lambda_1||_*||z_i||$ and limit the domain as a compact set. We have $\max_{||z||=\alpha}(-\frac{1}{2}(1+y_i)(w+\lambda_1)^T z) = \alpha||w+\lambda_1||_*$. So it yields that

$$g = \begin{cases} 1 - y_i(w^T x_i + b) & ||w+\lambda_1||_* \le \lambda_2 \\ +\infty \quad \text{otherwise} \end{cases}.$$

Similarly,

$$\max_{z_i}[-\lambda_1^T z_i - \lambda_2||z_i||] = \begin{cases} 0 & ||\lambda_1||_* \leq \lambda_2 \\ +\infty & \text{otherwise} \end{cases}.$$

Minimize the dual function can obtain the dual problem as follows,

$$\begin{aligned} &\min_{\lambda_1,\lambda_2} g(\lambda_1,\lambda_2) = \min_{\lambda_1,\lambda_2} (1 - y_i(w^T x_i + b))^+ + \lambda_1^T m_i + \lambda_2 \sigma_i \\ &s.t. \quad ||w + \lambda_1||_* \leq \lambda_2, ||\lambda_1||_* \leq \lambda_2. \end{aligned} \tag{16}$$

The equivalent expression can be obtained by introducing constraints into objective function,

$$\min_{\lambda_1,\lambda_2} g(\lambda_1,\lambda_2) = (1 - y_i(w^T x_i + b))^+ + \min_{\lambda_1}\{\lambda_1^T m_i + \sigma_i \max[||\lambda_1 + w||_*, ||\lambda_1||_*\}.$$

Since $||w + \lambda_1||_* \geq ||w||_* - ||\lambda_1||_*$, we have the lower bound

$$\min_{\lambda_1,\lambda_2} g(\lambda_1,\lambda_2) \geq (1 - y_i(w^T x_i + b))^+ + \min_{\lambda_1}\{\lambda_1^T m_i + \sigma_i \max[||w||_* - ||\lambda_1||_*, ||\lambda_1||_*\}.$$

There are two cases: 1) $||w||_* - ||\lambda_1||_* \geq ||\lambda_1||_*$, it follows that $\min_{||\lambda_1||_* \leq \frac{1}{2}||w||_*}\{\lambda_1^T m_i + \sigma_i(||w||_* - ||\lambda_1||_*)\}$ has the optimal lower bound

$$\min_{\lambda_1,\lambda_2} g(\lambda_1,\lambda_2) \geq (1 - y_i(w^T x_i + b))^+ + \frac{1}{2}(\sigma_i - ||m_i||)||w||_*.$$

And, 2) $||w||_* - ||\lambda_1||_* \leq ||\lambda_1||_*$ means the optimal lower bound of $\min_{||\lambda_1||_* \geq \frac{1}{2}||w||_*}\{\lambda_1^T m_i + \sigma_i||\lambda_1||_*\}$ is

$$\min_{\lambda_1,\lambda_2} g(\lambda_1,\lambda_2) \geq (1 - y_i(w^T x_i + b))^+ + \frac{1}{2}(\sigma_i - ||m_i||)||w||_*.$$

The lower bound is achieved when $\lambda_1 = -\frac{1}{2}w$. Thus the formulation with similar SVM form is given by

$$\min \frac{1}{2}(\sigma_i - ||m_i||)||w||_* + (1 - y_i(w^T x_i + b))^T + \frac{u^T u}{1^T s}. \tag{17}$$

Here, we take the expectation in the dual norm term for simplicity and randomized the SVM classifier in (17) to consider the approximate problem,

$$\begin{aligned} &\min \frac{1}{2}\sum_i (\sigma_i - ||m_i||)||u||_* + \frac{u^T u}{1^T s} \\ &s.t. \ P_{w\sim N(u,\Sigma)}(a_i^T w \leq -1 - t_i) \geq \nu, i = 1, 2, ...n. \end{aligned} \tag{18}$$

The $a_i = -y_i x_i$ makes the above constraint the same as (5). Rewrite it by Gauss distribution and introduce slack variable, the final formulation yields,

$$\begin{aligned} &\min \frac{1}{2}\sum_i (\sigma_i - ||m_i||)||u||_* + \frac{u^T u}{1^T s} + C\sum_i \xi_i \\ &s.t. \ \xi_i \geq 0, i = 1, 2, ..., n, \qquad s_i \geq 0, i = 1, 2, ..., d, \\ &\qquad y_i^T(u^T x_i) \geq 1 + \Phi^{-1}(\nu)\sum_{j=1}^{d} x_{ij}^2 s_j - \xi_i. \end{aligned} \tag{19}$$

6 Experimental Evaluation

In this section, we evaluate our proposed randomized SVM models against different type of attacks and compare their performance with the robust SVM model proposed in [13]. When simulating different types of attacks, we should estimate x_{ij}^t, the target to which the adversary may change x_i (note that the actual modification is $x_i + \delta_i$, where δ_i depends on x_{ij}^t). We use a simple method to estimate $x_{ij}^t = x_{ij} + c_\delta \varepsilon_{ij}$, where $c_\delta \in (0, 1)$ controls the aggressiveness of attacks (c_δ is small if attackers are conservative on modifying positive samples within a small area, while c_δ is large if attackers are aggressive on poisoning a larger range of sample space), and $\varepsilon_{ij} \sim \mathcal{N}(0, 1)$.

The attack strategies are depicted by setting δ_i in different attack models defined in (1) for RRA, (3) for RRAU, and (4) for DRAU, where c_f in (1) and (3), and σ in (4) control the intensity of attacks (how much to modify x_i). We will evaluate the proposed randomized SVM models at different levels of aggressiveness and intensity. In addition, the matrix P in RRAU is set to be a diagonal matrix, whose maximum eigenvalues $eig(P) \leq 0.5$. In attack model DRAU, we set the parameter σ differently and let $m = 0$. Attacks with larger σ are intenser with higher uncertainty.

We have three different randomized SVM models designed for various attacks, RRA, RRAU and DRAU. Since they are all randomized SVM, we differentiate them by the attack model names, such as SVM-RRA, SVM-RRAU and SVM-DRAU. In SVM-RRAU, we set $||.||_*$ as L_2 norm and $||v_i||$ as L_1 norm in its optimization function (13). In SVM-DRAU, different norms L_1, L_2 and L_∞ are studied in (19).

Four binary classification data sets (SEEDS, CLIMATE, QSAR, and SPAMBASE) from UCI repository are used as evaluation data. Linear SVMs were implemented using the LIBSVM library, while CVX package with SDPT3.0, MOSEK, Gurobi solvers is implemented to solve randomized classifier SVM models against different attacks. The probability of separation ν in (5) is set to 0.59 if not especially specified.

The experiments go through the following steps for obtaining performance measurement:

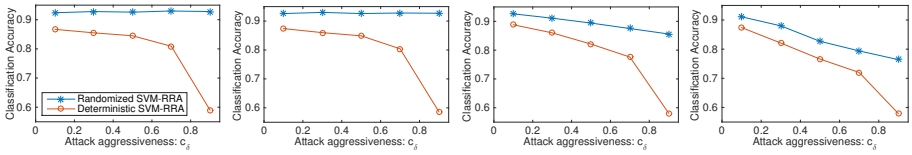
1). Load the training data $(x_i, y_i), i = 1, ..., n$,

2). Modify each x_i to obtain x_i^t,

3). Train classifier distribution $\mathcal{N}_w(u, s * s^T)$,

4). Obtain $y_i^t = sign(w^T x_i + b)$ to evaluate classification accuracy and failure rate.

Accuracy measures the classification correctness: Accuracy $= \frac{\sum_i \{y_i^t \neq y_i\}}{n}$. A high value of accuracy indicates the strong capability of the classification model to differentiate one class from the other. Failure rate measures how much the classification system fails to resist the attacks. Failure rate $= \frac{\sum_i \{y_i^t = -1 | y_i = 1\}}{n}$, i.e., among the manipulated malicious data samples with $y_i = 1$, how much of them are recognized as innocuous (the system fails and the attacker wins).



(a) $c_f = 0.3$ (b) $c_f = 0.5$ (c) $c_f = 0.7$ (d) $c_f = 0.9$

Fig. 2: Classification accuracy of randomized SVM-RRA and deterministic SVM-RRA when training data are poisoned at different levels of attack aggressiveness and intensity.

Table 1: Accuracy of SVM-RRAU when varying c_f, c_δ

	$c_f \downarrow$	c_δ=0.1	c_δ=0.3	c_δ=0.5	c_δ=0.7	c_δ=0.9
SVM-RRAU	0.1	0.92	0.92	0.91	0.90	0.90
	0.3	0.92	0.90	0.90	0.88	0.86
	0.5	0.91	0.89	0.87	0.84	0.82
	0.7	0.91	0.88	0.84	0.81	0.80
	0.9	0.90	0.87	0.81	0.80	0.80
Standard SVM		c_δ=0.1	c_δ=0.3	c_δ=0.5	c_δ=0.7	c_δ=0.9
		0.86	0.71	0.65	0.62	0.62

Table 2: Accuracy of SVM-DRAU with L_∞ when varying σ, ν

	$\sigma \downarrow$	ν=0.59	ν=0.69	ν=0.79	ν=0.89	ν=0.99
SVM-DRAU with L_∞	0.1	0.91	0.91	0.91	0.91	0.91
	0.3	0.90	0.90	0.90	0.89	0.89
	0.5	0.89	0.89	0.89	0.89	0.88
	0.7	0.88	0.88	0.88	0.87	0.87
	0.9	0.88	0.88	0.87	0.87	0.87
Standard SVM		σ=0.1	σ=0.3	σ=0.5	σ=0.7	σ=0.9
		0.86	0.71	0.65	0.62	0.62

6.1 Comparison with Deterministic SVM-RRA

We compare our **randomized** SVM-RRA with the **deterministic** SVM-RRA in [13]. and show how randomization improves the robustness of SVM against RR attack. Figure 2 shows the comparison of classification accuracy when applying **randomized** SVM-RRA and **deterministic** SVM-RRA on poisoned training data at different levels of attack aggressiveness and intensity. The evaluation is on SPAMBASE data set. In each subfigure of Figure 2, the attack aggressiveness (x-axis) varied from least ($c_\delta = 0.1$) to most aggressive ($c_\delta = 0.9$). The attack intensity varies from gentle ($c_f = 0.3$) to intensest ($c_f = 0.9$).

The overall conclusion we can draw from Figure 2 is that our randomized SVM-RRA always has higher accuracy than the deterministic SVM-RRA in [13]. The advantage of using randomization for enhancing the robustness of classification system is significant. When attacks are gentle ($c_f = 0.3$ and 0.5), our randomized SVM-RRA performs well with a stable high accuracy even when attackers are aggressive on attacking a large region. However, the performance of deterministic SVM-RRA decreases significantly when c_δ increases. When attacks are intense ($c_f \geq 0.7$), the accuracy of both our randomized SVM-RRA

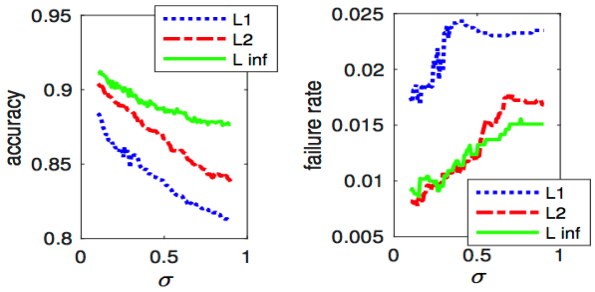


Fig. 3: Accuracy (upper part) and failure rate (lower part) of SVM-DRAU with different norms when varying σ.

and deterministic SVM-RRA is affected by aggressive attacks. However, our randomized SVM-RRA always performs better than the deterministic SVM-RRA.

6.2 Comparative Study with Standard SVMs

Attacks with uncertainty are less studied in literature. We evaluate the performance of our randomized SVM-RRAU and SVM-DRAU with L_∞ against attacks at different level of intensity and aggressiveness, compared to the performance of standard SVMs. In SVM-RRAU, the intensity and aggressiveness are controlled by c_f and c_δ respectively. In SVM-DRAU with L_∞, we vary ν in (5) to see under different separation probability how SVM-DRAU with L_∞ performs in classification against attacks at different level of aggressiveness (controlled by σ). The results given in Table 1 and 2 are averaged on four UCI data sets.

There are several observations from these tables. First, all randomized SVM models are more robust against generalized attacks, comparing to standard SVM. Second, model robustness decreases when attacks are more aggressive and more severe. Third, in Table 2, when ν increases, the classifier is expected to separate more examples with a high probability. Then the accuracy decreases a little as it probably makes wrong separations in order to meet the requirement of ν.

6.3 SVM-DRAU with different norms

To further analyze the performance of SVM-DRAU with different norms, we evaluate the *classification accuracy* and *resistance failure rate* at different level of attack intensity. Here we fix the separation probability ν in (5) to be 0.59 and set attack aggressiveness $c_\delta = 0.3$. When changing the attack intensity σ from 0.1 to 0.9, the performance of SVM-DRAU with different norms is shown in Figure 3. We can see that obviously SVM-DRAU with L_∞ is the most accurate and most robust one. It always has higher accuracy than others, while keeps resistance failure rate low. SVM-DRAU with L_2 also has low failure rate, but lower accuracy than SVM-DRAU with L_∞.

7 Conclusion

In this paper, we investigate how randomization can improve the robustness of SVMs against attack models with uncertainty. We define two general attack models and design randomized SVM models for each attack model. The randomized SVMs are formulated as standard convex optimization problems. Experimental results illustrate the effectiveness of our proposed models on several datasets and their better performance than baseline methods.

Acknowledgments

This work was supported by the King Abdullah University of Science and Technology, and by Natural Science Foundation of China, under grant U1736114 and 61672092, and in part by National Key R&D Program of China (2017YFB0802805).

References

1. Alabdulmohsin, I.M., Gao, X., Zhang, X.: Adding robustness to support vector machines against adversarial reverse engineering. In: CIKM. pp. 231–240 (2014)
2. Barreno, M., Nelson, B., Sears, R., Joseph, A.D., Tygar, J.D.: Can machine learning be secure? In: Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security. pp. 16–25 (2006)
3. Bi, J., Zhang, T.: Support vector classification with input data uncertainty. In: NIPS. pp. 161–168 (2004)
4. Biggio, B., Corona, I., Nelson, B., Rubinstein, B.I.P., Maiorca, D., Fumera, G., Giacinto, G., Roli, F.: Security evaluation of support vector machines in adversarial environments. Support Vector Machines Applications pp. 105–153 (2014)
5. Biggio, B., Nelson, B., Laskov, P.: Support vector machines under adversarial label noise. In: Proceedings of the 3rd Asian Conference on Machine Learning. pp. 97–112 (2011)
6. Brückner, M., Kanzow, C., Scheffer, T.: Static prediction games for adversarial learning problems. Journal of Machine Learning Research **13**(1), 2617–2654 (Sep 2012)
7. Dalvi, N., Domingos, P., Mausam, Sanghai, S., Verma, D.: Adversarial classification. In: SIGKDD. pp. 99–108 (2004)
8. Dekel, O., Shamir, O., Xiao, L.: Learning to classify with missing and corrupted features. Machine Learning **81**(2), 149–178 (2010)
9. Globerson, A., Roweis, S.: Nightmare at test time: Robust learning by feature deletion. In: ICML. pp. 353–360 (2006)
10. Großhans, M., Sawade, C., Brückner, M., Scheffer, T.: Bayesian games for adversarial regression problems. In: ICML. pp. 55–63 (2013)
11. Xu, H., Caramanis, C., Mannor, S.: Robustness and regularization of support vector machines. Journal of Machine Learning Research **10**, 1485–1510 (Dec 2009)
12. Xu, L., Crammer, K., Schuurmans, D.: Robust support vector machine training via convex outlier ablation. In: AAAI. pp. 536–542 (2006)
13. Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Xi, B.: Adversarial support vector machine learning. In: SIGKDD. pp. 1059–1067 (2012)