Advocating SVMs Against Poisoning Attacks: The Hardness and DBSCAN Approach

Hu Ding, Fan Yang, Jiawei Huang

1School of Computer Science and Technology, University of Science and Technology of China, He Fei, China

Abstract

Adversarial machine learning has attracted a great amount of attention in recent years. Due to the great importance of support vector machines (SVM) in machine learning, we consider defending SVM against poisoning attacks in this paper. We study two commonly used strategies for defending: designing robust SVM algorithms and data sanitization. Though several robust SVM algorithms have been proposed before, most of them either are in lack of adversarial-resilience, or rely on strong assumptions about the data distribution or the attacker’s behavior. Moreover, the research on the hardness of designing a quality-guaranteed adversarially-resilient SVM algorithm is still quite limited. We are the first, to the best of our knowledge, to prove that even the simplest hard-margin one-class SVM with adversarial outliers problem is NP-complete, and has no fully PTAS unless \( P=NP \). For data sanitization, we explain the effectiveness of DBSCAN (as a density-based outlier removal method) for defending against poisoning attacks. In particular, we link it to the intrinsic dimensionality by proving a sampling theorem in doubling metrics. In our empirical experiments, we systematically compare several defenses including the DBSCAN and robust SVM methods, and investigate the influences from the intrinsic dimensionality and poisoned fraction to their performances.

1 INTRODUCTION

In the past decades we have witnessed enormous progress in machine learning. One driving force behind this is the successful applications of machine learning technologies to many different fields, such as data mining, networking, and bioinformatics. However, with its territory rapidly enlarging, machine learning has also imposed a number of new challenges. In particular, adversarial machine learning which concerns about the potential vulnerabilities of the algorithms, has attracted a great amount of attention [Barreno et al., 2006, Huang et al., 2011, Biggio and Roli, 2018, Goodfellow et al., 2018]. As mentioned in the survey paper [Biggio and Roli, 2018], the very first work of adversarial machine learning dates back to 2004, in which Dalvi et al. [2004] formulated the adversarial classification problem as a game between the classifier and the adversary. In general, the adversarial attacks against machine learning can be categorized to evasion attacks and poisoning attacks [Biggio and Roli, 2018]. An evasion attack happens at test time, where the adversary aims to evade the trained classifier by manipulating test examples. For example, Szegedy et al. [2014] observed that small perturbation to a test image can arbitrarily change the neural network’s prediction.

In this paper, we focus on poisoning attacks that happen at training time. Usually, the adversary injects a small number of specially crafted samples into the training data which can make the decision boundary severely deviate; in particular, because open datasets are commonly used to train our machine learning algorithms nowadays, poisoning attack has become a key security issue that seriously limits real-world applications [Biggio and Roli, 2018]. For instance, even a small number of poisoning samples can significantly increase the test error of support vector machine (SVM) [Biggio et al., 2012, Mei and Zhu, 2015, Xiao et al., 2012]. Beyond linear classifiers, a number of works studied the poisoning attacks for other machine learning problems, such as clustering [Biggio et al., 2013], PCA [Rubinstein et al., 2009], and regression [Jagielski et al., 2018].

Though lots of works focused on constructing poisoning attacks, our ultimate goal is to design defenses. Poisoning samples can be regarded as outliers, and this leads to two natural approaches to defend: (1) data sanitization defense, i.e., first perform outlier removal and then run an existing machine learning algorithm on the cleaned data [Cretu et al., 2008], or (2) directly design a robust optimization algo-
ritment that is resilient against outliers [Christmann and Steinwart 2004; Jagielski et al. 2018].

Steinhardt et al. [2017] studied two basic methods of data sanitization defense, which remove the points outside a specified sphere or slab, for binary classification; they showed that high dimensionality gives attacker more room for constructing attacks to evade outlier removal. Laishram and Phoha [2016] applied the seminal DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method [Ester et al. 1996] to remove outliers for SVM and showed that it can successfully identify most of the poisoning data. However, their DBSCAN approach is lacking of theoretical analysis. Several other outlier removal methods for fighting poisoning attacks have also been studied recently [Paudice et al. 2018b]. Also, it is worth noting that outlier removal actually is an independent topic that has been extensively studied in various fields before [Chandola et al. 2009].

The other defense strategy, designing robust optimization algorithms, also has a long history in the machine learning community. A substantial part of robust optimization algorithms rely on the idea of regularization. For example, Xu et al. [2009] studied the relation between robustness and regularization for SVM; Zhu et al. [2003] proposed the 1-norm SVM to enhance the robustness to noise; other robust SVM algorithms include [Tax and Duin 1999; Xu et al. 2006; Natarajan et al. 2013; Ding and Xu 2015; Xu et al. 2017; Kanamori et al. 2017]. However, as discussed in [Mei and Zhu 2015; Jagielski et al. 2018], these approaches are not quite ideal to defend against poisoning attacks since the outliers can be located arbitrarily in the feature space by the adversary. Another idea for achieving the robustness guarantee is to add strong assumptions about the data distribution or the attacker’s behavior [Feng et al. 2014; Weerasinghe et al. 2019], but these assumptions are usually not well satisfied in practice. An alternative approach is to explicitly remove outliers during optimization, such as the “trimmed” method for robust regression [Jagielski et al. 2018]; but this is often a challenging combinatorial optimization problem: if \( z \) of the input \( n \) data items are outliers \( (z < n) \), (at first glance) we have to consider an exponentially large number \( \binom{n}{z} \) of different possible cases in the adversarial setting.

1.1 OUR CONTRIBUTIONS

Due to the great importance in machine learning [Chang and Lin 2011], we focus on defending SVM against poisoning attacks in this paper. Our contributions are twofold.

(i). First, we consider the robust optimization approach. To study its complexity, we only consider the hard-margin case (because the soft-margin case is more complicated and thus should have an even higher complexity). As mentioned above, we can formulate the SVM with outliers problem as a combinatorial optimization problem for achieving the adversarial-resilience: finding an optimal subset of \( n - z \) items from the poisoned input data to achieve the largest separating margin. Though its local optimum can be obtained by using various methods, such as the alternating minimization approach [Jagielski et al. 2018], it is very challenging to achieve a quality guaranteed solution for such adversarial-resilience optimization problem. For instance, Simonov et al. [2019] showed that unless the Exponential Time Hypothesis (ETH) fails, it is impossible not only to solve the PCA with outliers problem exactly but even to approximate it within a constant factor. A similar hardness result was also proved for linear regression with outliers by Mount et al. [2014]. Some other hardness results for robust optimization problems were studied in [Bernholt 2006]. But for SVM with outliers, we are unaware of any hardness-of-approximation result before. We try to bridge the gap in the current state of knowledge in Section 3. We prove that even the simplest one-class SVM with outliers problem is NP-complete, and has no fully polynomial-time approximation scheme (PTAS) unless P=NP. So it is quite unlikely that one can achieve a (nearly) optimal solution in polynomial time.

(ii). Second, we investigate the DBSCAN based data sanitization defense and explain its effectiveness in theory (Section 4). DBSCAN is one of the most popular density-based clustering methods and has been implemented for solving many real-world outlier removal problems [Ester et al. 1996; Schubert et al. 2017]; roughly speaking, the outliers are assumed to be located in some dense regions and the remaining points are recognized as the outliers. Actually, the intuition of using DBSCAN for data sanitization is straightforward [Laishram and Phoha 2016]. We assume the original input training data (before poisoning attack) is large and dense enough in the domain \( \Omega \); thus the poisoning data should be the sparse outliers together with some small clusters located outside the dense regions, which can be identified by the DBSCAN. Obviously, if the attacker has a fixed budget \( z \) (the number of poisoning points), the larger the data size \( n \) is, the sparser the outliers appear to be (and the more efficiently the DBSCAN performs).

Thus, to guarantee the effectiveness of the DBSCAN approach, a fundamental question in theory is what about the lower bound of the data size \( n \) (we can assume that the original input data is a set of i.i.d. samples drawn from the domain \( \Omega \)). However, to achieve a favorable lower bound is a non-trivial task. The VC dimension [Li et al. 2001] of the range space induced by the Euclidean distance is high in a high-dimensional feature space, and thus the lower bound of the data size \( n \) can be very large. Our idea is motivated by the recent observations on the link between the adversarial vulnerability and the intrinsic dimensionality [Khoury and Hadfield-Menell 2019; Amsaleg et al. 2017; Ma et al. 2018]. We prove a lower bound of \( n \) that depends on the
intrinsic dimension of $\Omega$ and is independent of the feature space’s dimensionality.

Our result strengthens the observation from Steinhardt et al. [2017] who only considered the Euclidean space’s dimensionality: more precisely, it is the “high intrinsic dimensionality” that gives attacker more room to evade outlier removal. In particular, different from the previous results on evasion attacks [Khoury and Hadfield-Menell 2019, Amsaleg et al. 2017, Ma et al. 2018], our result links poisoning attacks to intrinsic dimensionality (independent of our work, Weerasinghe et al. [2021] recently also studied the relation between intrinsic dimension and poisoning attacks). In Section 5, we investigate several popular defending methods (including DBSCAN), where the intrinsic dimension of data demonstrates significant influence on their defending performances.

2 PRELIMINARIES

Given two point sets $P^+$ and $P^-$ in $\mathbb{R}^d$, the problem of linear support vector machine (SVM) [Chang and Lin 2011] is to find the maximum margin separating these two point sets (if they are separable). If $P^+$ (or $P^-$) is a single point, say the origin, the problem is called one-class SVM. The SVM can be formulated as a quadratic programming problem, and a number of efficient techniques have been developed in the past, such as the soft margin SVM [Cortes and Vapnik 1995], $\nu$-SVM [Scholkopf et al. 2000, Crisp and Burges 1999], and Core-SVM [Tsang et al. 2005].

We consider two point sets $P$ and $O$, and suppose that $P^+ \cup P^- = n$. Since poisoning attack is expensive, we often assume that the adversary can poison at most $z \in \mathbb{Z}^+$ points (or the poisoned fraction $\frac{z}{n}$ is a fixed small number in $[0, 1]$). We can formulate the defense against poisoning attacks as the following combinatorial optimization problem. As mentioned in Section 1.1, it is sufficient to consider only the simpler hard-margin case for studying its hardness.

Definition 1 (SVM with Outliers). Let $(P^+, P^-)$ be an instance of SVM in $\mathbb{R}^d$, and suppose $|P^+ \cup P^-| = n$. Given a positive integer $z < n$, the problem of SVM with outliers is to find two subsets $P_1^+ \subseteq P^+$ and $P_1^- \subseteq P^-$ with $|P_1^+ \cup P_1^-| = n - z$, such that the width of the margin (i.e., the distance between the two parallel hyperplanes bounding the margin) separating $P_1^+$ and $P_1^-$ is maximized.

Suppose the optimal margin has the width $h_{opt} > 0$. If we achieve a solution with the margin width $h \geq (1 - \epsilon)h_{opt}$, where $\epsilon$ is a small number in $(0, 1)$, we say that it is a $(1 - \epsilon)$-approximation.

Remark 1. The model proposed in Definition 1 follows the popular data trimming idea from robust statistics [Rousseeuw and Leroy 1987]. As an example similar with Definition 1, Jagielski et al. [2018] proposed a data trimming based regression model to defend against poisoning attacks.

We also need to clarify the intrinsic dimensionality for our following analysis. Doubling dimension is a measure of intrinsic dimensionality that has been widely adopted in the learning theory community [Bshouty et al. 2009]. Given a point $p$ and $r \geq 0$, we use $B(p, r)$ to indicate the ball of radius $r$ around $p$ in the space.

Definition 2 (Doubling Dimension). The doubling dimension of a point set $P$ from some metric space $\mathbb{X}$ is the smallest number $\rho$, such that for any $p \in P$ and $r \geq 0$, the set $P \cap B(p, 2r)$ can always be covered by the union of at most $2^\rho$ balls with radius $r$ in the space.

To understand doubling dimension, we consider the following simple case. If the points of $P$ distribute uniformly in a $d'$-dimensional flat in $\mathbb{R}^d$, then it is easy to see that $P$ has the doubling dimension $\rho = O(d')$, which is independent of the Euclidean dimension $d$ (e.g., $d$ can be much higher than $\rho$). Intuitively, doubling dimension is used for describing the expansion rate of a given point set in the space. It is worth noting that the intrinsic dimensionality described in [Amsaleg et al. 2017, Ma et al. 2018] is quite similar to doubling dimension, which also measures expansion rate.

3 THE HARDNESS RESULT

In this section, we prove that even the one-class SVM with outliers problem is NP-complete and has no fully PTAS unless P=NP (that is, we cannot achieve a polynomial time $(1 - \epsilon)$-approximation for any given $\epsilon \in (0, 1)$). Our idea is partly inspired by the result from Megiddo [1990]. Given a set of points in $\mathbb{R}^d$, the “covering by two balls” problem is to determine that whether the point set can be covered by two unit balls. By the reduction from 3-SAT, Megiddo proved that the “covering by two balls” problem is NP-complete. In the proof of the following theorem, we modify Megiddo’s construction of the reduction to adapt the one-class SVM with outliers problem.

Theorem 1. The one-class SVM with outliers problem is NP-complete, and has no fully PTAS unless P=NP.
Let \( \Gamma \) be a 3-SAT instance with the literal set \( \{u_1, \bar{u}_1, \cdots, u_l, \bar{u}_l\} \) and clause set \( \{E_1, \cdots, E_m\} \). We construct the corresponding instance \( P_\Gamma \) of one-class SVM with outliers. First, let \( U = \{\pm e_i \mid i = 1, 2, \cdots, l + 1\} \) be the \( 2(l + 1) \) unit vectors of \( \mathbb{R}^{l+1} \), where each \( e_i \) has “1” in the \( i \)-th position and “0” in other positions. Also, for each clause \( E_j \) with \( 1 \leq j \leq m \), we generate a point \( q_j = (q_{j,1}, q_{j,2}, \cdots, q_{j,l+1}) \) as follows. For \( 1 \leq i \leq l \),

\[
q_{j,i} = \begin{cases} 
\alpha, & \text{if } u_i \text{ occurs in } E_j; \\
-\alpha, & \text{if } \bar{u}_i \text{ occurs in } E_j; \\
0, & \text{otherwise.}
\end{cases}
\]

In addition, \( q_{j,l+1} = 3\alpha \). For example, if \( E_j = u_{i_1} \lor \bar{u}_{i_2} \lor u_{i_3} \), the point

\[
q_j = (0, \cdots, 0, \alpha, 0, \cdots, 0, -\alpha, 0, \cdots, 0, 0, 0, \cdots, 0, 3\alpha).
\]

The value of \( \alpha \) will be determined later. Let \( Q \) denote the set \( \{q_1, \cdots, q_m\} \). Now, we construct the instance \( P_\Gamma = U \cup Q \) of one-class SVM with outliers, where the number of points \( n = 2(l + 1) + m \) and the number of outliers \( z = l + 1 \). Then we have the following lemma.

**Lemma 1.** Let \( \alpha > 1/2 \), \( \Gamma \) has a satisfying assignment if and only if \( P_\Gamma \) has a solution with margin width \( \frac{1}{\sqrt{l+1}} \).

**Proof.** First, we suppose there exists a satisfying assignment \( \mathcal{A}(\Gamma) \) for \( \Gamma \). We define the set \( S \subset P_\Gamma \) as follows. If \( u_i \) is true in \( \mathcal{A}(\Gamma) \), we include \( e_i \) in \( S \), else, we include \( -e_i \) in \( S \). We also include \( e_{l+1} \) in \( S \). We claim that the set \( S \cup Q \) yields a solution of the instance \( P_\Gamma \) with the margin width \( \frac{1}{\sqrt{l+1}} \), that is, the size \( |S \cup Q| = n - z \) and the margin separating the origin \( o \) and \( S \cup Q \) has width \( \frac{1}{\sqrt{l+1}} \). It is easy to verify the size of \( S \cup Q \). To compute the width, we consider the mean point of \( S \) which is denoted as \( t \). For each \( 1 \leq i \leq l \), if \( u_i \) is true, the \( i \)-th position of \( t \) should be \( \frac{1}{\sqrt{l+1}} \), else, the \( i \)-th position of \( t \) should be \( -\frac{1}{\sqrt{l+1}} \), the \((l+1)\)-th position of \( t \) is \( \frac{1}{\sqrt{l+1}} \). Let \( \mathcal{H}_t \) be the hyperplane that is orthogonal to the vector \( t - o \) and passing through \( t \). It is easy to know \( \mathcal{H}_t \) separates \( S \) and \( o \), and the margin width (i.e., the distance between the origin and \( \mathcal{H}_t \)) is \( ||t|| = \frac{1}{\sqrt{l+1}} \). Furthermore, for any point \( q_j \in Q \), since there exists at least one true variable in \( E_j \), we have the inner product

\[
\langle q_j, \frac{t}{||t||} \rangle \geq \frac{3\alpha}{\sqrt{l+1}} + \frac{\alpha}{\sqrt{l+1}} - \frac{2\alpha}{\sqrt{l+1}} = \frac{2\alpha}{\sqrt{l+1}} > 1
\]

where the last inequality comes from the fact \( \alpha > 1/2 \). Therefore, all the points from \( Q \) lie on the same side of \( \mathcal{H}_t \) as \( S \), and then the set \( S \cup Q \) can be separated from \( o \) by a margin with width \( \frac{1}{\sqrt{l+1}} \).

Second, suppose the instance \( P_\Gamma \) has a solution with margin width \( \frac{1}{\sqrt{l+1}} \). With a slight abuse of notations, we still use \( S \) to denote the subset of \( U \) that is included in the set of \( n - z \) inliers. Since the number of outliers is \( z = l + 1 \), we know that for any pair \( \pm e_i \), there exists exactly one point belonging to \( S \); also, the whole set \( Q \) should be included in the set of inliers so as to guarantee that there are \( n - z \) inliers in total. We still use \( t \) to denote the mean point of \( S \) \( (||t|| = \frac{1}{\sqrt{l+1}}) \). Now, we design the assignment \( \mathcal{A}(\Gamma) \) for \( \Gamma \): if \( e_i \in S \), we assign \( u_i \) to be true, else, we assign \( \bar{u}_i \) to be false. We claim that \( \Gamma \) is satisfied by this assignment. For any clause \( E_j \), if it is not satisfied, i.e., all the three variables in \( E_j \) are false, then we have the inner product

\[
\langle q_j, \frac{t}{||t||} \rangle \leq \frac{3\alpha}{\sqrt{l+1}} - \frac{3\alpha}{\sqrt{l+1}} = 0.
\]

That means the angle \( \angle q_j, ot \geq \pi/2 \). So any margin separating the origin \( o \) and the set \( S \cup Q \) should have the width at most

\[
\frac{||q_i|| \cdot ||t||}{\sqrt{||q_i||^2 + ||t||^2}} < ||t|| = \frac{1}{\sqrt{l+1}}.
\]

See Figure 1 for an illustration. This is in contradiction to the assumption that \( P_\Gamma \) has a solution with margin width \( \frac{1}{\sqrt{l+1}} \).

Overall, \( \Gamma \) has a satisfying assignment if and only if \( P_\Gamma \) has a solution with margin width \( \frac{1}{\sqrt{l+1}} \). \( \square \)

Now we are ready to prove the theorem.

**Proof.** (of Theorem 1) Since 3-SAT is NP-complete, Lemma 1 implies that the one-class SVM with outliers problem is NP-complete too; otherwise, we can determine that
whether a given instance $\Gamma$ is satisfiable by computing the optimal solution of $P_1$. Moreover, the gap between \( \sqrt{\frac{1}{\sqrt{m}}} \) and \( \frac{1}{\sqrt{m}} \) (from the formula 44) is

\[
\frac{1}{\sqrt{m}+\sqrt{\frac{1}{\sqrt{m}}}} = \Theta\left(\frac{1}{\sqrt{m}}\right),
\]

if we assume $\alpha$ is a fixed constant. Therefore, if we set $\epsilon = O\left(\frac{\alpha}{\sqrt{m}}\right) = O\left(\frac{1}{\sqrt{m}}\right)$, then $\Gamma$ is satisfiable if and only if any $(1-\epsilon)$-approximation of the instance $P_1$ has width $\geq \sqrt{\frac{12n^2 + 12n^2 + 2\sqrt{3}\alpha}{\sqrt{m} + \sqrt{\frac{1}{\sqrt{m}}}}}$. That means if we have a fully PTAS for the one-class SVM with outliers problem, we can determine whether $\Gamma$ is satisfiable or not in polynomial time. In other words, we cannot even achieve a fully PTAS for one-class SVM with outliers, unless $P=NP$.

4 THE DATA SANITIZATION DEFENSE

From Theorem 1, we know that it is extremely challenging to achieve the optimal solution even for one-class SVM with outliers. Therefore, we turn to consider the other approach, data sanitization defense, under some reasonable assumption in practice. First, we prove a general sampling theorem in Section 4.1. Then, we apply this theorem to explain the effectiveness of DBSCAN for defending against poisoning attacks in Section 4.2.

4.1 A SAMPLING THEOREM

Let $P$ be a set of i.i.d. samples drawn from a connected and compact domain $\Omega$ who has the doubling dimension $\rho > 0$. For ease of presentation, we assume that $\Omega$ lies on a manifold $\mathcal{F}$ in the space. Let $\Delta$ denote the diameter of $\Omega$, i.e., $\Delta = \sup_{p_1, p_2 \in \Omega} ||p_1 - p_2||$. Also, we let $f$ be the probability density function of the data distribution over $\Omega$.

To measure the uniformity of $f$, we define a weight $\lambda$ as follows. For any $c \in \Omega$ and any $r > 0$, we say “the ball $B(c, r)$ is enclosed by $\Omega$” if $\partial B(c, r) \cap \Omega \subset \Omega$; intuitively, if the ball center $c$ is close to the boundary $\partial \Omega$ of $\Omega$ or the radius $r$ is too large, the ball will not be enclosed by $\Omega$. See Figure 1b for an illustration. We define $\lambda := \sup_{c \in \Omega} \int_{B(c,r)} f(x) \, dx$, where $B(c, r)$ and $B(c', r)$ are any two equal-sized balls, and $B(c, r)$ is required to be enclosed by $\Omega$. As a simple example, if $\Omega$ lies on a flat manifold and the data uniformly distribute over $\Omega$, the value $\lambda$ will be equal to 1. On the other hand, if the distribution is very imbalanced or the manifold $\mathcal{F}$ is very rugged, the value $\lambda$ can be high.

Theorem 2. Let $m \in \mathbb{Z}^+$, $\epsilon \in (0, \frac{1}{8})$, and $\delta \in (0, \Delta)$. If the sample size

\[
|P| > \max\left\{ \Theta\left(\frac{m - \epsilon}{1 - \epsilon} \cdot \lambda \cdot \left(\frac{1 + \epsilon \Delta}{1 - \epsilon \delta}\right)^\rho\right), \hat{O}(\rho \cdot \lambda^2 \cdot \left(\frac{1 + \epsilon \Delta}{1 - \epsilon \delta}\right)^{2\rho}\left(\frac{1}{\epsilon}\right)^{\rho+2}) \right\},
\]

then with constant probability, for any ball $B(c, \delta)$ enclosed by $\Omega$, the size $|B(c, \delta) \cap P| > m$. The asymptotic notation $\hat{O}(f) = \Theta(f \cdot \text{polylog}(\frac{1}{\delta}))$.

Remark 2. (i) A highlight of Theorem 2 is that the lower bound of $|P|$ is independent of the dimensionality of the input space (which could be much higher than the intrinsic dimension). Moreover, our result holds for any metric space with bounded doubling dimension (not only for Euclidean space).

(ii) For the simplest case that $\Omega$ lies on a flat manifold and the data uniformly distribute over $\Omega$, $\lambda$ will be equal to 1 and thus the lower bound of $|P|$ in Theorem 2 becomes

\[
\max\left\{ \Theta\left(\frac{m - \epsilon}{1 - \epsilon} \cdot \left(\frac{1 + \epsilon \Delta}{1 - \epsilon \delta}\right)^\rho\right), \hat{O}(\rho \cdot \left(\frac{1 + \epsilon \Delta}{1 - \epsilon \delta}\right)^{2\rho}\left(\frac{1}{\epsilon}\right)^{\rho+2}) \right\}.
\]

Before proving Theorem 2, we need to relate the doubling dimension $\rho$ to the VC dimension $\mathfrak{dim}$ of the range space consisting of all balls with different radii \cite{Li2001}. Unfortunately, \cite{Huang2018} recently showed that “although both dimensions are subjects of extensive research, to the best of our knowledge, there is no nontrivial relation known between the two”. For instance, they constructed a doubling metric having unbounded VC dimension, and the other direction cannot be bounded neither. However, if allowing a small distortion to the distance, we can achieve an upper bound on the VC dimension for a given metric space with bounded doubling dimension. For stating the result, they defined a distance function called “$\epsilon$-smoothed distance function”: $g(p, q) \in (1 \pm \epsilon)||p - q||$ for any two data points $p$ and $q$, where $\epsilon \in (0, \frac{1}{4})$. Given a point $p$ and $\delta > 0$, the ball defined by this distance function $g(\cdot, \cdot)$ is denoted by $B_\delta(p, \delta) = \{q \in \text{input space} \mid g(p, q) \leq \delta\}$.

Theorem 3 (Huang et al. 2018). Suppose the point set $P$ has the doubling dimension $\rho > 0$. There exists an $\epsilon$-smoothed distance function “$g(\cdot, \cdot)$” such that the VC dimension $\mathfrak{dim}$ of the range space consisting of all balls with different radii is at most $O(\frac{1}{\epsilon^2})$. If replacing the distance by $g(\cdot, \cdot)$.

Proof. (of Theorem 2) Let $r$ be any positive number. First, since the doubling dimension of $\Omega$ is $\rho$, if recursively applying Definition 2 $\log \frac{\Delta}{r}$ times, we know that $\Omega$ can be divided into $O(\frac{\Delta}{r})$ parts. Hence, we have $\hat{O}(\rho_0 \log \rho_0)$.
covered by at most $\Theta\left(\left(\frac{r}{\lambda}\right)^\rho\right)$ balls with radius $r$. Thus, if \( B(c, r) \) is enclosed by $\Omega$, we have

$$
\frac{\int_{B(c,r)} f(x) \, dx}{\int_{\Omega} f(x) \, dx} \geq \Theta\left(\frac{1}{\lambda} \cdot \left(\frac{r}{\Delta}\right)^\rho\right). \tag{7}
$$

Now we consider the size $|B(c, \delta) \cap \Omega|$. From Theorem 3, we know that the VC dimension $\dim r \Theta$ with constant probability $\epsilon$, (7), (9), and (10) jointly imply

$$
|P| \geq \Theta\left(\frac{1}{\epsilon}\dim r \log \frac{\dim r}{\epsilon}\right), \tag{8}
$$

the set $P$ will be an $\epsilon_0$-sample of $\Omega$; that is, for any point $c$ and $\delta' \geq 0$,

$$
\frac{|B_g(c, \delta') \cap P|}{|P|} \leq \frac{\int_{B_g(c,\delta')} f(x) \, dx}{\int_{\Omega} f(x) \, dx} \pm \epsilon_0 \tag{9}
$$

with constant probability $\epsilon_0$. Because $g(\cdot, \cdot)$ is an $\epsilon$-smoothed distance function of the Euclidean distance, we have

$$
B(c, \frac{\delta'}{1+\epsilon}) \subseteq B_g(c, \delta') \subseteq B(c, \frac{\delta'}{1-\epsilon}). \tag{10}
$$

So if we set $\epsilon_0 = \epsilon \cdot \Theta\left(\frac{1}{\lambda} \cdot \left(\frac{1-\epsilon}{1+\epsilon} \frac{\delta'}{1-\epsilon}\right)^\rho\right)$ and $\delta' = (1-\epsilon)\delta$, (7), (9), and (10) jointly imply

$$
\frac{|B(c, \frac{\delta'}{1+\epsilon}) \cap P|}{|P|} \geq \frac{|B_g(c, \delta') \cap P|}{|P|} \geq \frac{\int_{B_g(c,\delta')} f(x) \, dx}{\int_{\Omega} f(x) \, dx} - \epsilon_0 \geq \frac{\int_{B(c, \frac{\delta'}{1+\epsilon})} f(x) \, dx}{\int_{\Omega} f(x) \, dx} - \epsilon_0 \geq \frac{1}{\lambda} \cdot \Theta\left(\frac{1-\epsilon}{1-\epsilon} \frac{\delta'}{1-\epsilon}\right)^\rho. \tag{11}
$$

The last inequality comes from (7) (since we assume the ball $B(c, \delta)$ is enclosed by $\Omega$, the shrunk ball $B(c, \frac{\delta'}{1+\epsilon}) = B(c, \frac{\delta'}{1+\epsilon})$ should be enclosed as well). Moreover, if

$$
|P| \geq \Theta\left(\frac{m}{1-\epsilon} \cdot \lambda \cdot \left(1 + \epsilon \frac{\Delta}{1-\epsilon}\right)^\rho\right), \tag{12}
$$

we have $|B(c, \delta) \cap P| > m$ from (11). Combining (8) and (12), we obtain the lower bound of $|P|$. \( \square \)

4.2 THE DBSCAN APPROACH

For the sake of completeness, we briefly introduce the method of DBSCAN [Ester et al. 1996]. Given two parameters $r > 0$ and $\text{MinPts} \in \mathbb{Z}^+$, the DBSCAN divides the set $P$ into three classes: (1) $p$ is a core point, if $|B(p, r) \cap P| > \text{MinPts}$; (2) $p$ is a border point, if $p$ is not a core point but $p \in B(q, r)$ of some core point $q$; (3) all the other points are outliers. Actually, we can imagine that the set $P$ forms a graph where any pair of core or border points are connected if their pairwise distance is no larger than $r$; then the set of core points and border points form several clusters where each cluster is a connected component (a border point may belong to multiple clusters, but we can arbitrarily assign it to only one cluster). The goal of DBSCAN is to identify these clusters and the outliers. Several efficient implementations for DBSCAN can be found in [Gan and Tao 2013, Schubert et al. 2017]. Following Section 4.1, we assume that $P$ is a set of i.i.d. samples drawn from the connected and compact domain $\Omega$ who has the doubling dimension $\rho > 0$. We let $Q$ be the set of $z$ poisoning data items injected by the attacker to $P$, and suppose each $q \in Q$ has distance larger than $\delta_1 > 0$ to $\Omega$. In an evasion attack, we often use the adversarial perturbation distance to evaluate the attacker’s capability; but in a poisoning attack, the attacker can easily achieve a large perturbation distance (e.g., in the SVM problem, if the attacker flips the label of some point $p$, it will become an outlier having the perturbation distance larger than $h_{opt}$ to its ground truth domain, where $h_{opt}$ is the optimal margin width). Also, we assume the boundary $\partial \Omega$ is smooth and has curvature radius at least $\delta_2 > 0$ everywhere. For simplicity, let $\delta = \min\{\delta_1, \delta_2\}$. The following theorem states the effectiveness of the DBSCAN with respect to the poisoned dataset $P' \cup Q$. We assume the poisoned fraction $|Q|/|P'| = 1/Q' < 1$.

**Theorem 4.** We let $m$ be any absolute constant number larger than 1, and assume that the size of $P$ satisfies the lower bound of Theorem 2. If we set $r = \delta$ and $\text{MinPts} = m$, and run DBSCAN on the poisoned dataset $P' \cup Q$, then the obtained largest cluster is exactly the set $P$. In other word, the set $Q$ consists of the outliers and the clusters except the largest one from the DBSCAN.

**Proof.** Since $\delta \leq \delta_2$, for any $p \in P$, either the ball $B(p, \delta)$ is enclosed by $\Omega$, or $p$ is covered by some ball $B(q, \delta)$ enclosed by $\Omega$. We set $r = \delta$ and $\text{MinPts} = m$, and hence from Theorem 2 we know that all the points of $P$ will be core points or border points. Moreover, any point $q$ from $Q$ has distance larger than $r$ to the points of $P$, that is, any two points $q \in Q$ and $p \in P$ should not belong to the same cluster of the DBSCAN. Also, because the domain $\Omega$ is connected and compact, the set $P$ must form the largest cluster. \( \square \)
Remark 3. (i) We often adopt the poisoned fraction $\frac{S}{m}$ as the measure to indicate the attacker’s capability. If we fix the value of $z$, the bound of $|P|$ from Theorem 2 reveals that the larger the doubling dimension $\rho$, the lower the poisoned fraction $\frac{S}{m}$ (and the easier corrupting the DBSCAN defense). In addition, when $\delta$ is large, i.e., each poisoning point has large perturbation distance and $\delta$ is sufficiently smooth, it will be relatively easy for DBSCAN to defend.

But we should point out that this theoretical bound probably is overly conservative, since it requires a “perfect” sanitization result that removes all the poisoning samples (this is not always a necessary condition for achieving a good defending performance in practice). In our experiments, we show that the DBSCAN method can achieve promising performance, even when the poisoned fraction is higher than the threshold.

(ii) In practice, we cannot obtain the exact values of $\delta$ and $m$. We follow the strategy that was commonly used in the DBSCAN implementations [Gan and Tao 2015, Schubert et al. 2017]: we set $\text{MinPts}$ to be a small constant and tune the value of $r$ until the largest cluster has $|P \cup Q| - z$ points.

Putting it all together. Let $(P^+, P^-)$ be an instance of SVM with $z$ outliers, where $z$ is the number of poisoning points. We assume that the original input point sets $P^+$ and $P^-$ (before the poisoning attack) are i.i.d. samples drawn respectively from the connected and compact domains $\Omega^+$ and $\Omega^-$ with doubling dimension $\rho$. Then, we perform the DBSCAN procedure on $P^+$ and $P^-$ respectively (as Remark 3(ii)). Suppose the obtained largest clusters are $P^+$ and $P^-$. Finally, we run an existing SVM algorithm on the cleaned instance $(\hat{P}^+, \hat{P}^-)$.

5 EMPIRICAL EXPERIMENTS

All the experiments were repeated 20 times on a Windows 10 workstation equipped with an Intel core i5-8400 processor and 8GB RAM. To generate the poisoning attacks, we use the MIN-MAX attack from [Koh et al. 2018] and the adversarial label-flipping attack ALFA from ALFASVMLib [Xiao et al. 2015]. We evaluate the defending performances of the basic SVM algorithms and several different defenses by using their publicly available implementations.

1. We consider both the cases that not using and using kernel. For SVM without kernel, we directly use LINEAR SVM as the basic SVM algorithm; for SVM with kernel, we consider RBF kernel (RBF SVM). Both the implementations are from [Chang and Lin 2011].

2. The recently proposed robust SVM algorithm RSVM-S based on the rescaled hinge loss function [Xu et al. 2017]. The parameter “$S$” indicates the iteration number of the half-quadratic optimization (e.g., we set $S = 3$ and 10 following their paper’s setting). The algorithm also works fine when using a kernel.

3. The DBSCAN method [Schubert et al. 2017] implemented as Remark 3(ii). We set $\text{MinPts} = 5$ (our empirical study finds that the difference is minor within the range [3, 10]).

4. The data sanitization defenses from [Koh et al. 2018] based on the spatial distribution of input data, which include SLAB, L2, LOSS, and K-NN.

For the data sanitization defenses, we run them on the poisoned data in the original input space; then, apply the basic SVM algorithm, LINEAR SVM or RBF SVM (if using RBF kernel), on the cleaned data to compute their final solutions.

Table 1: Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYNTHETIC</td>
<td>10000</td>
<td>50-200</td>
</tr>
<tr>
<td>LETTER</td>
<td>1520</td>
<td>16</td>
</tr>
<tr>
<td>MUSHROOMS</td>
<td>8124</td>
<td>112</td>
</tr>
<tr>
<td>SATIMAGE</td>
<td>2236</td>
<td>36</td>
</tr>
</tbody>
</table>

Datasets. We consider both the synthetic and real-world datasets in our experiments. For each synthetic dataset, we generate two manifolds in $\mathbb{R}^d$, and each manifold is represented by a random polynomial function with degree $d'$ (the values of $d$ and $d'$ will be varied in the experiments). Note that it is challenging to achieve the exact doubling dimensions of the datasets, and thus we use the degree of the polynomial function as a “rough indicator” for the doubling dimension (the higher the degree, the larger the doubling dimension). In each of the manifolds, we randomly sample 5000 points; the data is randomly partitioned into 30% and 70% respectively for training and testing, and we report the classification accuracy on the test data. We also consider three real-world datasets from [Chang and Lin 2011]. The details are shown in Table 1.

Results. First, we study the influence from the intrinsic dimensionality. We set the Euclidean dimensionality $d$ to be 100 and vary the polynomial function’s degree $d'$ from 25 to 65 in Figure 2a and 2d. Then, we fix the degree $d'$ to be 40 and vary the Euclidean dimensionality $d$ in Figure 2b and 2e. We can observe that the accuracies of most methods dramatically decrease when the degree $d'$ (intrinsic dimension) increases, and the influence from the intrinsic dimension is more significant than that from the Euclidean dimension. We also study their classification performances under different poisoned fraction in Figure 2f and 2i. We can see that all the defenses yield lower accuracies when the poisoned fraction increases, while the performance of DBSCAN keeps much more stable compared with other defenses. Moreover, we calculate the widely used $F_1$ scores from the sanitization
Figure 2: The classification accuracy on the SYNTHETIC datasets of Linear SVM (the first line) and SVM with RBF kernel (the second line) under MIN-MAX attack. The third line are the average $F_1$ scores.

Figure 3: The classification accuracy on the real datasets of linear SVM (the first line) and SVM with RBF kernel (the second line) under MIN-MAX attack.
we also perform the experiments on the real datasets under M\textsuperscript{in-Max} attack and ALFA attack with the poisoned fraction ranging from 4\% to 10\%. The experimental results (Figure 3 and 4) reveal the similar trends as the results for the synthetic datasets, and DBSCAN keeps considerably better performance compared with other defenses. The \( F_1 \) scores on MUSHROOM dataset are shown in Table 2 (due to the space limit, the \( F_1 \) scores on the other two real datasets are placed in our full paper).

We also perform the experiments on the real datasets under M\textsuperscript{in-Max} attack and ALFA attack with the poisoned fraction ranging from 4\% to 10\%. The experimental results (Figure 3 and 4) reveal the similar trends as the results for the synthetic datasets, and DBSCAN keeps considerably better performance compared with other defenses. The \( F_1 \) scores on MUSHROOM dataset are shown in Table 2 (due to the space limit, the \( F_1 \) scores on the other two real datasets are placed in our full paper).

6 DISCUSSION

In this paper, we study two different strategies for protecting SVM against poisoning attacks. We also have several open questions to study in future. For example, what about the complexities of other machine learning problems under the adversarially-resilient formulations as Definition 1? For many other adversarial machine learning problems, the study on their complexities is still in its infancy.

Acknowledgements

The authors would like to thank Ruomin Huang and the anonymous reviewers for their helpful discussions and suggestions on improving this paper.

References


