Unsupervised Adversarial Anomaly Detection using One-Class Support Vector Machines

(Extended Abstract)

Prameesha S. Weerasinghe[1], Sarah M. Erfani[2], Tansu Alpcan[1], Christopher Leckie[2] and Margreta Kuijper[1]

Abstract— **Anomaly detection discovers regular patterns in unlabeled data and identifies the non-conforming data points, which in some cases are the result of malicious attacks by adversaries. Learners such as One-Class Support Vector Machines (OCSVMs) have been successfully used in anomaly detection, yet their performance may degrade significantly in adversarial conditions such as integrity attacks. This work focuses on integrity attacks, where the adversary distorts the training data in order to successfully avoid detection during evaluation. This paper presents a unique combination of anomaly detection using (1) OCSVMs in the presence of adversaries who distort training data in a targeted manner and (2) nonlinear randomized kernel methods, which facilitate computational and conceptual simplification through dimension reduction. We theoretically analyze the effects of adversarial distortions on the separating margin of OCSVMs and provide supporting empirical evidence. The proposed approach introduces a layer of uncertainty on top of the OCSVM learner, making it challenging for the adversary to guess the specific configuration of the learner.**

Index Terms— **Machine learning, unsupervised learning, adversarial learning.**

I. INTRODUCTION

Machine learning algorithms have been successfully employed in a wide range of domains such as finance, health, security etc [1]. The basis of many machine learning algorithms is to learn the underlying hidden structure and patterns from data, without being explicitly programmed. Once the patterns are learned from the data, the algorithm can apply this knowledge on unseen samples with significant accuracy.

Anomaly detection is a class of machine learning that is associated with the problem of discovering patterns in data and identifying data points that do not conform to the learned patterns (i.e., outliers). It has numerous applications in a variety of domains such as network intrusion detection, credit card fraud detection, and spam filtering. Algorithms such as One-Class Support Vector Machines (OCSVM) [2], have been proven to be effective in anomaly detection applications. Although they are designed to withstand the effects of random noise in data, when adversaries deliberately alter the input data, the performance of these learning algorithms may degrade significantly.

Anomaly detection systems are often deployed in environments where the data naturally evolves, and hence the models need to be retrained periodically, in contrast to many conventional machine learning applications, where the current and future data is assumed to have identical properties. This periodic training may allow adversaries to gradually inject malicious data to diminish the decision making capabilities of the learning algorithms. The aim of the adversaries may be to reduce the risk of being detected (i.e., attack on integrity) or to decrease the performance of the learning system [3].

A sophisticated adversary has the capacity to conduct an attack in numerous ways. Hence, it is not feasible to provide a general analysis that covers the whole range of attacks, across different machine learning algorithms. In this work, we explore the following key question: *Is it possible to make OCSVMs more resistant against adversarial attacks which target the integrity of the training data through distortions?*. If an adversary can maliciously distort the input data used by a learning algorithm, they can force the learner to learn a model that is favorable to them. It has become imperative to secure machine learning systems against such adversaries due to the recent increase of automation in many day to day applications.

For example, in the context of image recognition, the adversary could cause distortions that are imperceptible to humans, but influential enough to force a learned model to mis-classify the distorted images with high confidence. As [4] have shown, with the emergence of self driving vehicles, an adversary could alter a *"S-T-O-P"* road sign in such a way that a vehicle (learning system) would reliably classify it as a *"Speed Limit 45"* sign. Such distortions could be imperceptible to humans and could result in the loss of human lives.

To mitigate this issue, we utilize a nonlinear data projection based algorithm to increase the attack resistance of OCSVMs against an adversarial opponent under realistic assumptions. Recent work in the literature shows that nonlinear random projections improve the training and evaluation times of kernel machines, without significantly compromising the accuracy of the trained models [5], [6]. In this paper, we investigate whether selective nonlinear random projections can be leveraged to increase the attack resistance of OCSVMs under adversarial conditions.

The main **contributions** of this work are summarized as follows. We propose a nonlinear data transformation based defence mechanism that can (i) increase the attack resistance of OCSVMs under adversarial conditions, and (ii) give the

[1] Authors are with the Electrical & Electronic Engineering Department, University of Melbourne, Parkville 3010 Victoria, Australia pweerasinghe@student.unimelb.edu.au, tansu.alpcan@unimelb.edu.au, mkuijper@unimelb.edu.au

[2] Authors are with the School of Computing and Information Systems, University of Melbourne, Parkville 3010 Victoria, Australia caleckie@unimelb.edu.au, sarah.erfani@unimelb.edu.au

learner a significant advantage from a security perspective by adding a layer of unpredictability through the randomness of the data transformation. The adversary's goal is to hinder the decision making capabilities of the OCSVM by shifting its decision boundary using distorted data points. It should be noted that the learner cannot demarcate adversarial distortions from the normal data, otherwise the learner would be able to remove the adversarial distortions during training, making the problem trivial. Due to this reason, the margin of a OCSVM trained only on clean data cannot be calculated empirically. Therefore, we analytically derive an upper bound on the length of the weight vector of a OCSVM trained on an undistorted dataset that has been nonlinearly transformed to a lower dimensional space.

II. BACKGROUND AND RELATED WORK

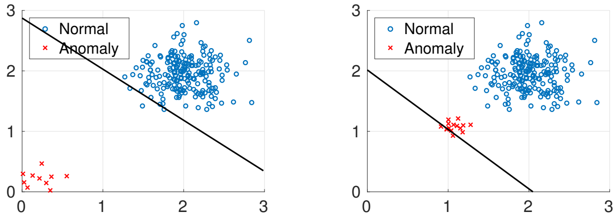
As our proposed approach on adversarial learning for anomaly detection is based on randomized kernels, in this section we briefly review these two lines of research.

To improve the efficiency of kernel machines, in [5], Rahimi and Recht embedded a random projection into the kernel formulation. They introduced a novel, data independent method (Random Kitchen Sinks (RKS)) that approximates a kernel function by mapping the dataset to a relatively low dimensional randomized feature space. Instead of relying on the implicit transformation provided by the kernel trick, they explicitly mapped the data to a low-dimensional Euclidean inner product space using a randomized feature map $z : \mathbb{R}^d \to \mathbb{R}^r$.

More recently, the method of [5] has been applied to other types of kernel machines. In [6], Erfani et al. introduced *Randomized One-class SVMs (R1SVM)*, an unsupervised anomaly detection technique that uses randomized, nonlinear features in conjunction with a linear kernel. They reported that R1SVM reduces the training and evaluation times of OCSVMs by up to two orders of magnitude without compromising the accuracy of the predictor. Our work differs from these as we look at random projections as a defense mechanism for OCSVMs under adversarial conditions. However, to the best of our knowledge, no existing work adopts Rahimi and Recht's method to address adversarial learning for anomaly detection with OCSVMs.

The problem of adversarial learning has inspired a wide range of research from the machine learning community, see [7] for a survey. For example, [8] introduced an Adversarial SVM (AD-SVM) model. AD-SVM incorporated additional constraint conditions to the binary SVM optimization problem in order to thwart an adversary's attacks. Their model leads to unsatisfactory results when the severity of real attacks differs from the model's expected attack severity. While we gain valuable insights regarding attack strategies from this work, the defense mechanism in our work is significantly different and our work primarily focuses on unsupervised learning, whereas [8] uses a binary SVM.

Deep Neural Networks (DNNs) have been shown to be robust to noise in the input [9], but are unable to withstand carefully crafted adversarial data [10]. While these works are in the same domain, they are not directly related to our work, which uses OCSVMs and kernels.



(a) no attack

(b) attack on integrity

Fig. 1: Training data distribution and separating hyperplane (black line) of a toy problem with and without an attack. ‘o’ (blue) denotes the undistorted data points and ‘x’ (red) denotes the data points distorted by the adversary. The OCSVM is trained using the entire (unlabeled) dataset as normal.

III. PROBLEM STATEMENT AND ATTACK MODEL

We consider an adversarial learning problem for anomaly detection in the presence of a malicious adversary. The adversary modifies the training data in order to disrupt the learning process of the learner, who aims to detect anomalous data points. Hence, the adversary's main goal is to hinder the decision making capability of the learner by compromising the integrity of the input data.

In an *integrity attack*, the adversary desires false negatives (i.e., anomalies classified as normal), and hence, would use distorted anomalies during training to move the decision boundary of the learner away from the normal data cloud and towards the anomalies. Subsequently, during the testing phase, any anomalies that lie beyond the compromised decision boundary will be classified as normal data points. As Figure 1a depicts, in the context of OCSVMs, the decision boundary (i.e., separating hyperplane) is found closer to the normal data cloud. The adversary would distort anomalies in order to place them closer to the normal data cloud. Since the OCSVM algorithm considers all the data points in the training set to be from the normal class (i.e., only uses data from normal class during training), these distorted anomalies would be seen by the learning algorithm as normal data points (similar to label flipping). As Figure 1b depicts, this would result in the separating hyperplane moving closer to the origin. The adversary is able to orchestrate different attacks by changing the percentage of distorted anomaly data points in the training dataset (i.e., p_{attack}) and the severity of the distortion (i.e., s_{attack}).

The attack model used is inspired by the restrained attack model described by [8]. Let $X + D$ be the training dataset that contains the data from the normal class X as well as the adversarial distortions D. The adversary has the freedom to determine D based on the knowledge it possesses regarding the learning system, although the magnitude of D is usually bounded due to its limited knowledge about the learners' configuration, the increased risk of being discovered, and computational constraints. It is assumed that the adversary has the capability to move the i^{th} data point in any direction by adding a non-zero displacement vector $\kappa_i \in D$ to $x_i \in X$. It is also assumed that the adversary does not have any knowledge about the projection used by the learner. Therefore, all of the adversary's actions take place in the original full dimensional space. The adversary picks a target x_i^t for each x_i to be distorted and moves it towards the target by some amount. Choosing x_i^t for each x_i optimally requires a significant level of computational effort and a thorough knowledge about the distribution of the data. For each attribute j in the original feature space, the adversary is able to add κ_{ij} to x_{ij}, where

$$\kappa_{ij} = (1-s_{attack})(x_{ij}^t - x_{ij}) \text{ and } |\kappa_{ij}| \leq |x_{ij}^t - x_{ij}|, \forall j \in d. \quad (1)$$

IV. Methodology

Anticipating possible distortions by an adversary, the learner can take precautions to minimize their effects by contracting the data to a lower dimensional space. Projecting a high dimensional dataset, using a carefully chosen projection matrix would preserve its pairwise Euclidean distances with high probability in the projected space [11]. Therefore, the properties of the original data distribution would be present in the projected dataset with only minor perturbations. By randomly drawing projection directions from some distribution, the learner introduces a layer of uncertainty to the adversary-learner problem. For high dimensional datasets, this method gives the learner considerate flexibility to select the dimension to which the data is projected, as well as the direction, which gives a significant advantage from a security perspective. But this unpredictability can also be seen as the main caveat of using random projections to reduce the dimensionality of data. While some random projections result in better separated volumetric clouds than the original ones, some projections result in the data from different classes being overlapped.

In order to increase the attack resistance of a learning system, the impact of adversarial inputs should be minimized. Based on this intuition, we propose that a projection that conceals the potential distortions of an adversary would make any learning system that learns from the projected data more resistant to attacks. As the learner cannot demarcate D from the training data, it is not possible to identify an ideal projection that conceals the adversarial distortions. Thus, the learner would have to select a projection that contracts the entire training set (expecting the adversarial points to be masked by normal data) and separates the training data from the origin with the largest margin in the transformed space. Therefore, we propose a novel compactness measure to identify suitable projection directions in a one-class problem [12].

V. Impact of Attack on the OCSVM Margin

This section analyzes the effects of the adversary's distortions on the margin of separation of the OCSVM. The distance between the hyperplane and the origin of a OCSVM is given by $\rho/\|w\|_2$, where ρ is the offset and w is the vector of weights. This implies that a small $\|w\|_2$ corresponds to a large margin of separation from the origin. Since the learner cannot demarcate the distortions from the normal training data, it cannot empirically calculate this value for the undistorted dataset. Therefore, based on the assumptions given below, we analytically derive an upper bound on $\|w\|_2$ of a OCSVM that has been trained on a nonlinearly transformed undistorted dataset. This result would lead to a lower bound for the margin of separation of the OCSVM without any adversarial distortions. It should be noted that any attack on the integrity of the learner would be reflected on the margin of separation (i.e., a large change in the margin would indicate a successful attack). As the adversary distorts data in the original feature space, we can align any given dataset in such a way that any outliers present in the data would lie closer to the origin and the normal data cloud would lie in the positive orthant. Such a transformation would compel the adversary to make adversarial distortions in the direction of the normal data cloud (positive) as distortions in the negative direction would favor the learner.

Assumption 1. The distortions made by the adversary are small s.t. small angle approximation $\cos(\theta) = 1 - \frac{\theta^2}{2}$ holds.

This assumption is reasonable because small distortions decrease the risk of the adversary being discovered, therefore a rational adversary would refrain from conducting attacks with significant distortions.

Definition 1. Let $X \in \mathbb{R}^{n\times d}$ be the matrix that contains the training data (normalized between 0 – 1) and $D \in \mathbb{R}^{n\times d}$ the matrix that contains the adversarial distortions. Let $A \in \mathbb{R}^{d\times r}$ be the projection matrix where each element is an i.i.d. $\mathcal{N}(0,1)$ random variable. Define b as a $1\times r$ row vector where each element is drawn uniformly from $[0,2\pi]$. Using these variables, we define $C \in \mathbb{R}^{n\times r}$ (which is linearly separable [5]), where the element at row i column j takes the following form.

$$\begin{aligned} C_{i,j} = \cos\Big(\Big[&\big(X_{i,1}+D_{i,1}\big)A_{1,j} + \big(X_{i,2}+D_{i,2}\big)A_{2,j} + \dots \\ &+ \big(X_{i,d}+D_{i,d}\big)A_{d,j}\Big] + b_{1,j}\Big). \end{aligned} \quad (2)$$

Similarly, we define the matrices C^X, C^D, S^X, S^D as follows,

$$\begin{aligned} C^X_{i,j} &= \cos\Big(\big[X_{i,1}A_{1,j} + X_{i,2}A_{2,j} + \dots + X_{i,d}A_{d,j}\big] + b_{1,j}\Big), \\ C^D_{i,j} &= \cos\Big(\big[D_{i,1}A_{1,j} + D_{i,2}A_{2,j} + \dots + D_{i,d}A_{d,j}\big]\Big), \\ S^X_{i,j} &= \sin\Big(\big[X_{i,1}A_{1,j} + X_{i,2}A_{2,j} + \dots + X_{i,d}A_{d,j}\big] + b_{1,j}\Big), \\ S^D_{i,j} &= \sin\Big(\big[D_{i,1}A_{1,j} + D_{i,2}A_{2,j} + \dots + D_{i,d}A_{d,j}\big]\Big). \end{aligned}$$

We address the anomaly detection problem using the OCSVM algorithm introduced by [2], which separates the training data from the origin with a maximal margin in the transformed space. The dual form of the OCSVM algorithm can be written in matrix notation as,

$$\underset{\alpha}{\text{minimize}} \quad \frac{1}{2}\alpha^T CC^T \alpha, \ \text{s.t } 0 \leq \alpha \leq \frac{1}{\nu n} \text{ and } \mathbf{1}^T\alpha = 1, \quad (3)$$

where α is the vector of Lagrange multipliers, $\nu \in (0,1]$ is a parameter that defines an upper bound on the fraction of support vectors and a lower bound on the fraction of outliers, and $\mathbf{1}$ is a vector of ones.

Theorem 1. Let w_p^* be the primal solution of the OCSVM optimization problem in the transformed space without adversarial distortions. Similarly, define w_{pd}^* as the primal solution in the presence of a malicious adversary. Let r be the number of dimensions to which the data is transformed using the method in (2). Then, if Assumption 1 holds, the length of the weight vector w_p^* is bounded above by

$$\|w_p^*\|_2 \leq \|w_{pd}^*\|_2 + \frac{3\sqrt{r}}{2}. \qquad (4)$$

Proof outline: Let $\tilde{\alpha}$ be the vector achieving the optimal solution in the projected space when adversarial distortions are present. Then, the solution for the primal problem in the projected space with adversarial distortions, defined as w_{pd}^*, can be obtained as

$$\left\|w_{pd}^*\right\|_2 = \left\|\tilde{\alpha}^T C\right\|_2. \qquad (5)$$

Using the cosine angle-sum identity on the matrix defined by equation 2 (the symbol $\odot$ denotes the Hadamard product for matrices),

$$\left\|w_{pd}^*\right\|_2 = \left\|\tilde{\alpha}^T\left(C^X \odot C^D\right) - \tilde{\alpha}^T\left(S^X \odot S^D\right)\right\|_2. \qquad (6)$$

Using the reverse triangle inequality we obtain

$$\left\|w_{pd}^*\right\|_2 \geq \left\|\tilde{\alpha}^T\left(C^X \odot C^D\right)\right\|_2 - \left\|\tilde{\alpha}^T\left(S^X \odot S^D\right)\right\|_2. \qquad (7)$$

From the constraint conditions of the OCSVM problem (3), we get $\mathbf{1}^T\tilde{\alpha} = 1$. Also, as $\sin(\theta) \in [-1,1]$ the inequality can be further simplified as,

$$\left\|w_{pd}^*\right\|_2 \geq \left\|\tilde{\alpha}^T\left(C^X \odot C^D\right)\right\|_2 - \sqrt{r}. \qquad (8)$$

Due to *Assumption 1*, using small-angle approximation on C^D, followed by the reverse triangle inequality, we obtain

$$\left\|w_{pd}^*\right\|_2 \geq \left\|\tilde{\alpha}^T C^X\right\|_2 - \left\|\tilde{\alpha}^T\left(C^X \odot \left(\frac{DA \odot DA}{2}\right)\right)\right\|_2 - \sqrt{r}. \qquad (9)$$

As the training data is normalized between (0 – 1), the maximum distortion magnitude that can be achieved is 1. Also, as $\cos(\theta) \in [-1,1]$ and $\mathbf{1}^T\tilde{\alpha} = 1$, the inequality can be further simplified as,

$$\left\|w_{pd}^*\right\|_2 \geq \left\|\tilde{\alpha}^T C^X\right\|_2 - \frac{\sqrt{r}}{2} - \sqrt{r}. \qquad (10)$$

Since the optimization problem is a minimization problem, as shown in (3), the optimal solution for the OCSVM without any distortion (i.e., α^*) would give a value less than or equal to the value given by $\tilde{\alpha}$. Thus,

$$\left\|\alpha^{*,T} C^X\right\|_2 \leq \left\|w_{pd}^*\right\|_2 + \frac{3\sqrt{r}}{2}, \qquad (11)$$

$$\left\|w_p^*\right\|_2 \leq \left\|w_{pd}^*\right\|_2 + \frac{3\sqrt{r}}{2}. \qquad (12)$$

The strength of the adversary's attacks will be reflected on the value of upper bound and will increase with the strength of the attacks. The learner is able to make the upper bound of $\|w_p^*\|_2$ tighter by reducing the dimensionality of the dataset (i.e., r). Refer Table I for empirical validation that shows the consistency of the upper bound, which is about 6% higher than the empirical value for both dimensions.

TABLE I: Comparison of actual $\left\|w_p^*\right\|_2$, calculated on the MNIST data ($p_{attack} = 5\%$ and $s_{attack} = 0.5$) and the theoretical upperbound calculated using Theorem 1 in Section V.

# of dim	$\|w_p^*\|_2$	Upperbound	$\|w_p^*\|_2$ as % of upperbound
210	1,969.30	2,094.74	94.01%
393	2,734.40	2,908.03	94.03%

ACKNOWLEDGMENT

This work was supported in part by the Australian Research Council Discovery Project under Grant DP140100819 and by a grant from the Northrop Grumman Corporation.

REFERENCES

[1] I. Kononenko, “Machine Learning for Medical Diagnosis: History, State of the Art and Perspective,” *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, Aug. 2001.

[2] P. B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, “Support Vector Method for Novelty Detection,” in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 582–588.

[3] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, “Adversarial Machine Learning,” in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, 2011, pp. 43–58.

[4] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, “Robust Physical-World Attacks on Machine Learning Models,” *CoRR*, vol. abs/1707.08945, 2017.

[5] A. Rahimi and B. Recht, “Random Features for Large-Scale Kernel Machines,” in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 1177–1184.

[6] S. M. Erfani, M. Baktashmotlagh, S. Rajasegarar, S. Karunasekera, and C. Leckie, “R1SVM: A Randomised Nonlinear Approach to Large-Scale Anomaly Detection,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 432–438.

[7] M. Barreno, B. Nelson, A. D. Joseph, and J. Tygar, “The security of machine learning,” *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[8] Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and B. Xi, “Adversarial Support Vector Machine Learning,” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 1059–1067.

[9] A. Fawzi, S. Moosavi-Dezfooli, and P. Frossard, “Robustness of classifiers: from adversarial to random noise,” *CoRR*, vol. abs/1608.08967, 2016.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” *CoRR*, vol. abs/1412.6572, 2014.

[11] W. B. Johnson and J. Lindenstrauss, “Extensions of Lipschitz mappings into a Hilbert space,” *Contemporary mathematics*, vol. 26, no. 189-206, p. 1, 1984.

[12] P. S. Weerasinghe, S. M. Erfani, T. Alpcan, and C. Leckie, “Unsupervised Anomaly Detection Using One-class Support Vector Machines Under Attacks on Integrity,” *27th International Joint Conference on Artificial Intelligence*, in review.