Received: 25 January 2018 | Revised: 17 March 2018 | Accepted: 20 March 2018

DOI: 10.1002/widm.1259

WILEY WIREs DATA MINING AND KNOWLEDGE DISCOVERY

ADVANCED REVIEW

A survey of game theoretic approach for adversarial machine learning

对抗性机器学习中博弈论策略的研究

Yan Zhou[1] | Murat Kantarcioglu[1] | Bowei Xi[2]

[1]Department of Computer Science, University of Texas at Dallas, Richardson, Texas

[2]Department of Statistics, Purdue University, West Lafayette, Indiana

Correspondence
Murat Kantarcioglu, Department of Computer Science, University of Texas at Dallas, 2601 N. Floyd Road, Richardson, TX 75080-3021.
Email: muratk@utdallas.edu

Funding information
ARO, Grant/Award Number: W911NF-171-0356

前进，进步

The field of machine learning is progressing at a faster pace than ever before. Many organizations leverage machine learning tools to extract useful information from a massive amount of data. In particular, machine learning finds its application in cybersecurity that begins to enter the age of automation. However, machine learning applications in cybersecurity face unique challenges other domains rarely do—attacks from active adversaries. Problems in areas such as intrusion detection, banking fraud detection, spam filtering, and malware detection have to face challenges of adversarial attacks that modify data so that malicious instances would evade detection by the learning systems. The adversarial learning problem naturally resembles a game between the learning system and the adversary. In such a game, both players would attempt to play their best strategies against each other while maximizing their own payoffs. To solve the game, each player would search for an optimal strategy against the opponent based on the prediction of the opponent's strategy choice. The problem becomes even more complicated in settings where the learning system may have to deal with many adversaries of unknown types. Applying game-theoretic approach, robust learning techniques have been developed to specifically address adversarial attacks and the preliminary results are promising. In this review, we summarize these results.

This article is categorized under:
Technologies > Machine Learning
Fundamental Concepts of Data and Knowledge > Key Design Issues in Data Mining （数据挖掘）

KEYWORDS
adversarial machine learning, game theory

1 | INTRODUCTION

Machine learning is a methodology for discovering meaningful patterns from large amounts of data. Machine learning algorithms are specifically designed to analyze data from which the target concept is learned. If learning is supervised, it begins with a set of training data that has been labeled by domain experts. Learning proceeds by searching for a general concept that summarizes the data and outputs a predictive model that can predict a label for a new sample that is not in the training data. The reason that machine learning algorithms can learn from one set of training samples and make accurate predictions for other datasets lies in the fact that the training data and the future data have identical properties. Therefore, for machine learning algorithms to successfully learn the desired target concept, the independent and identically distributed (i.i.d.) assumption must hold—data used to train the learning system and the data on which the trained system is put to test are i.i.d. In other words, the set of training data is a faithful representation of the data in the entire domain. This i.i.d. assumption is the key to designing a traditional successful machine learning system.

违反 独立同分布假设

Unfortunately, cybersecurity applications typically have to face the unique challenge of violation of the i.i.d. assumption. For example, machine learning-based anti-virus (AV) systems are often challenged by malware authors who intentionally obfuscate malicious code to evade detection. The learning system (the defender) and the adversary (attackers as a whole) typically have opposed preferences, with each trying to defeat the other. The adversary attacks the learning system by strategically modifying malicious data so that the malicious data used in the training process are no longer i.i.d. samples from the malware data distribution encountered in the operating environment.

When adversarial attacks are performed on training data, the learning process is misguided, resulting in an inaccurate classifier. For example, the adversary may poison the training data by injecting strategically crafted samples to corrupt the learning process. Such attacks are referred to as poisoning attacks (Barreno et al., 2008; Biggio, Nelson, & Laskov, 2012). However, wielding direct influences on the learning process often requires privileges to manipulate training data and their labels, and therefore is more challenging for the adversary. On the other hand, disguising malicious instances at test time to evade detection by a trained classifier is easier and more practical (Biggio et al., 2013; Lowd & Meek, 2005; Wagner & Soto, 2002; Xu, Qi, & Evans, 2016). The second approach is more heavily researched in the field of adversarial machine learning. Barreno, Nelson, Joseph, and Tygar (2010) presented a detailed taxonomy of adversarial attacks.

Figure 1 illustrates an example of adversarial attacks against a trained classifier. On the left, we show a normal machine learning task of classifying samples into two classes, and the middle line is the decision boundary of the trained classifier. On the right, we show how modified malicious samples cross the decision boundary and get to be misclassified as benign instances. 经过修改的恶意样本如何跨越决策边界并被误分类为良性实例

The relationship between a learning system and an adversary resembles a two-player game. In many cybersecurity applications, the game seems never ending. Hence, the term “arms race” is often encountered in the related literature of adversarial machine learning.

To properly model such a learning problem as a two-player game, we have several issues to address. First, we have to decide upon the best way to model the game. We can model the problem as a strictly competitive game in which two players have diametrical preferences, meaning one player's gain is the other player's loss. This type of game is referred to as a zero sum game in which the total payoff is always zero. We can also model the game by assuming both players are only interested in maximizing their own payoffs where two players do not necessarily have completely opposed interests. Hence, we aim to search for an equilibrium state in which both players play their best strategies.

Besides the consideration of strict competitiveness, we also need to take into account the roles of the players in the game. We can model the game by assuming both players make their decisions simultaneously. On the other hand, there are certainly cases where one player may take the lead and the other player follows by playing the best strategy against the leader after knowing the leader's strategy choice. For example, spam filters are released as soon as e-mail service is provided. Spammers study a spam filter by sending messages for the spam filter to label. Other issues we have to address include whether the game stabilizes eventually and whether there exists equilibrium solutions for the game. In more realistic settings, a learning system may have to face many adversaries of several unknown types. In this type of adversarial learning problem, the learning system becomes the leader of multiple followers. The game is therefore referred to as a single leader multiple followers (SLMFs) game, which is a much more complicated case that demands additional care (Box 1).

一个领导者（learning system） 与 多个追随者（对抗性学习的敌手）之间的博弈，这是一种更为复杂的情况，需要额外关注

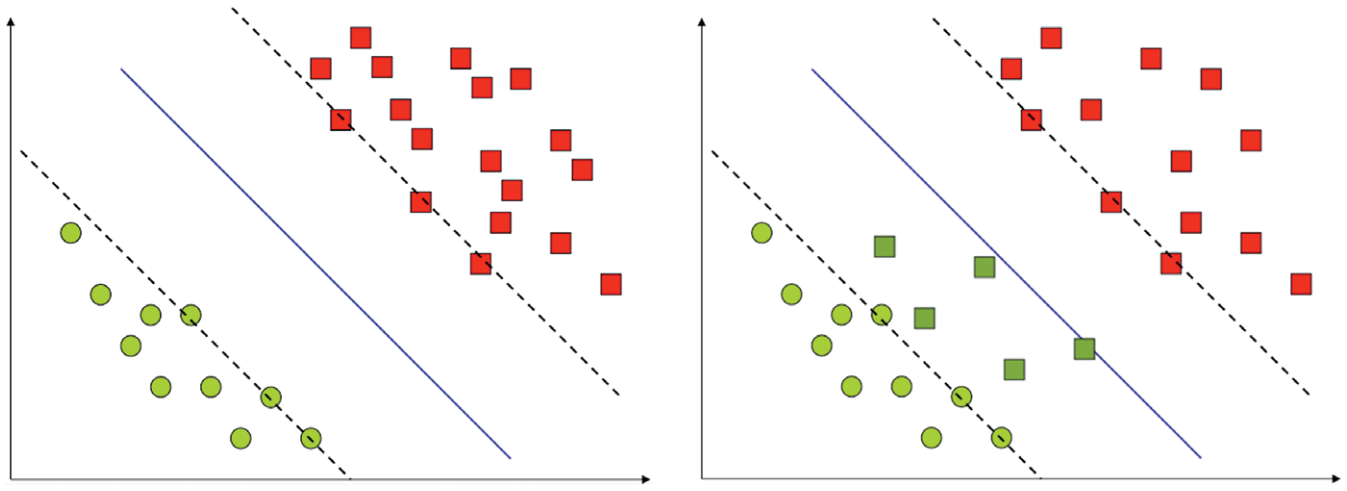

FIGURE 1 Adversarial attacks against a linear classifier on a two-dimensional dataset. The red squares represent malicious samples and the green dots are benign samples. The green squares are malicious samples modified by the adversary. The middle line is the decision boundary of the linear classifier

BOX 1

Real World Adversarial Attacks 以汽车为例，展示一个真实的 敌手对抗性攻击 的案例

With the adoption 使用 of a growing number of advanced technologies, modern automobiles have opened doors to adversarial attacks through their internal networked units. Recent research Koscher et al. (2010) has demonstrated that adversarial attacks can successfully circumvent driver's commands and perform adversarial controls such as disabling the brakes, 使刹车失灵 stopping the engine, permanently activating horn, turning windshield wipers on continuously, and many others. It has also been shown that it is possible to inject malicious code into a car's telematics unit to cause a crash 造成瘫痪 and erase all evidences of its presence afterwards. 事后，抹除所有痕迹

双重的 首先，对 对抗性机器学习 中用到的 博弈论模型 进行了一个综合性的 研究

The contribution of this paper is twofold. We first conduct a comprehensive survey of different game theoretic models used in adversarial learning. Then, we discuss the important insights gained from the game theoretic studies for building resilient learning algorithms against active adversaries. The paper is organized as follows. Section 2 introduces the game theoretic models. Section 3 discusses the important insights gained from the game theoretic studies. Section 4 concludes the paper.

从博弈论的角度来看 对抗性学习

2 | ADVERSARIAL LEARNING WITH A GAME THEORY PERSPECTIVE

The adversarial learning problem has been extensively studied under the game theoretic framework. Game theory provides a framework to study the strategic interactions between rational players. There are in general two different types of games applied to adversarial learning problem: *simultaneous* games and *sequential* games. In a simultaneous game, each player chooses its strategy without knowing the strategy its opponent is playing. In a sequential game, one player acts as the leader and chooses its strategy before the other players who later play optimally against the leader's strategy.

2.1 | Modeling adversarial learning as simultaneous games

Dalvi, Domingos, Mausam, and Verma (2004) consider an adversarial classification problem as a game between two cost-sensitive opponents—CLASSIFIER and ADVERSARY. For computational tractability, the classification game only considers move by each of the players. ADVERSARY is only allowed to modify positive instances. Both CLASSIFIER and ADVERSARY have a utility of classifying an instance of class y as y_C, denoted as $U_C(y_C, y)$ for CLASSIFIER and $U_A(y_C, y)$ for ADVERSARY, respectively. The goal of a player is to maximize the expected utility by playing an optimal strategy. ADVERSARY's optimal strategy is to transform a positive instance into a negative instance assuming it has complete information about CLASSIFIER and CLASSIFIER is unaware of its presence. On the other hand, CLASSIFIER assumes that ADVERSARY always plays its optimal strategies and the training set is not corrupted by ADVERSARY. The optimal strategy CLASSIFIER plays maximizes the postadversary conditional utility by predicting y_C, for an instance.

The adversarial learning problem is sometimes modeled as a zero-sum game between the learner and the adversary, in which one player's gain is the other player's loss. A zero-sum game is a strictly competitive game in that the two players' preferences are diametrically opposed and their expected payoffs always sum to zero. Therefore, maximizing one player's payoff is the same as minimizing the other player's payoff. In the game of adversarial learning, the learner searches for an optimal strategy that maximizes its expected payoff while anticipating the adversary is doing exactly the opposite, that is, minimizing the learner's expected payoff. The learner's problem of finding the optimal strategy is therefore the dual of the adversary's problem. The following discussions cover some related work in this direction:

- Globerson and Roweis (2006) consider a learning problem in which the adversary deletes features from samples in the test set. They formulate it as a minimax game for learning algorithms (e.g., support vector machine [SVM] Scholkopf and Smola [2001]) that are designed to minimize the hinge loss defined as $\max(0, 1 - yt)$, where t is the true class label and y is the predicted value (e.g., a probability) of a class. They set a feature of an instance to zero if it is deleted, and also specify a maximum number of features that can be deleted. They solve the optimization problem using quadratic programming. El Ghaoui, Lanckriet, and Natsoulis (2003) consider a similar problem in which they apply a minimax model to training data bounded by hyper-rectangles. They minimize the worst-case loss over samples in given intervals. There are several other variations of the adversarial learning problem regarding robust learning against classification-time noise (Dekel & Shamir, 2008; Dekel, Shamir, & Xiao, 2010; Lanckriet, Ghaoui, Bhattacharyya, & Jordan, 2002; Teo, Globerson, Roweis, & Smola, 2007).

支持向量机

- Our earlier work explores the impact of adversarial attack models on robust SVM (Zhou, Kantarcioglu, Thuraisingham, & Xi, 2012). We present an effective strategy for learning in the presence of adversaries that modify test data constrained with two attack models: *free-range attack* and *restrained attack*. The attack models are defined in terms of the adversary's capabilities of modifying data. Adversaries capable of free-range attacks have the freedom of modifying an instance to "move" it anywhere within the valid ranges in the feature space. This attack model is general enough to cover all possible attack scenarios in which test data is modified. In comparison, the restrained attack model is more realistic in that irrational data modifications are penalized. The adversary can only modify an instance so much that its malicious utility is not in jeopardy. In addition, the severity of attacks is controlled with a discount factor on data displacement. Our solutions minimize the expected loss corresponding to the two attack models. In general, learning an optimal classification strategy by assuming free-range adversarial attacks is often too pessimistic. When the real attacks are mild, the learning model that anticipates severe free-range attacks becomes too paranoid, and its classification accuracy decreases accordingly. On the contrary, learning with an assumption of restrained attacks produces outstanding overall performance regardless of how severe the real attacks are.
- Bruckner and Scheffer (2009) investigate whether there exists a unique Nash equilibrium for adversarial prediction games in which they assume both players commit to their strategies simultaneously. The uniqueness of a Nash equilibrium is important since it would be the only optimal pair of strategies for the learner and the adversary to play so that neither could benefit by unilaterally selecting a different strategy. In addition, both players must be rational in order to guarantee that the unique Nash equilibrium is optimal for both players. They prove that when the loss functions of both players are monotonic in the value of the classifier's prediction, that is, with one monotonically increasing and the other monotonically decreasing, the game has a unique Nash equilibrium and there exists an efficient algorithm to find it.

单调的

2.2 | Modeling adversarial learning as sequential games

Playing minimax strategies can become too pessimistic at times. On the other hand, in a sequential game, one player's strategy is exposed to the other players before they choose their strategies. The former is referred to as the leader of the game, and the latter is the follower. The leader must commit to a strategy before the follower responds. Once the leader commits to the strategy, it cannot be changed. The advantage the follower has is evident: a partial or complete knowledge of the leader. The follower can therefore play an optimal strategy against the leader. The leader, as the first mover, also has its own advantages. There are many successful examples of such first-mover advantages in real life. Similarly, in adversarial learning, the learner can leverage of first-mover advantages by controlling the cost the adversary has to pay in order to evade detection. For example, the learner can weigh the importance of features and make the most informative features that are indicative of malicious instances very difficult to modify without sacrificing the malicious utility. Such a game in which the first player is the leader and its opponent is the follower is known as a Stackelberg game. Research on modeling adversarial learning as Stackelberg games falls into two categories, depending on which players play the role of the leader: the learner or the adversary.

2.2.1 | Stackelberg games: Adversary as leader

Kantarcioglu, Xi, and Clifton (2011) consider the case where the classifier is the follower and can be improved after observing adversaries' strategies especially in the context of attribute selection. Attribute selection is an important aspect in constructing robust learning algorithms against adversaries. Attributes that are costly to modify will discourage adversaries to modify a very large number of malicious instances, and at the same time reduce the utility of malicious instances that successfully evade the detection. This work explores whether an equilibrium solution exists such that the learner's change of strategy to further reduce false negative cannot balance the increase in false positive; and vice versa, the adversary's change of strategy to transform a positive instance into a negative one cannot make up for the loss of malicious property. It is assumed that a player's utility function is known to the opponent. When both players behave rationally and play the Nash equilibrium strategies in every subgame, the pair of strategies is referred to as *subgame-perfect equilibrium*. Subgame perfect equilibrium is computed using simulated annealing and integer programming. The equilibrium is used to discover an optimal set of attributes to build the classifier.

Liu and Chawla (2010) also model the adversarial learning problem as a Stackelberg game in which the adversary is also the leader and the classifier is the follower. The game is defined as a maxmin optimization problem in which the adversary always plays the most rational strategy at each move. One difference in their research is that they relax the assumption that both players know their opponent's utility function. Instead, only the adversary's payoff function is required.

2.2.2 | Stackelberg games: Learner as leader

In cases where the learner needs to stay ahead of the game, the roles of the two players are normally switched. The learner can claim the role of leader in order to stay ahead of the curve and set the future races. For example, a machine learning-based

保持领先的曲线和设定未来的比赛

antivirus software is likely to commit to a strategy (after being trained on a carefully selected set of samples with high accuracy) before the adversary strikes back. The adversary, now in follower's seat with the knowledge about the learner's strategy, can play its best moves to defeat the learner. For example, sophisticated spammers would probably first probe spam filters to obtain firsthand knowledge about the spam classifiers before flooding e-mail servers with their spam messages. With a small set of carefully crafted e-mail messages, spammers can make an educated guess about where the decision boundaries are. Spammers empowered with the new knowledge can easily sabotage the first line defense of the spam filters.

Bruckner and Scheffer (2011) study this form of adversarial learning problems and present equilibrium solutions to the corresponding Stackelberg classification games. They define the adversary's cost as the sum of the expected prediction costs and the transformation costs and the learner's cost as the weighted prediction costs in which higher costs are paid for misclassified positive samples. The Stackelberg game is formalized as a bilevel optimization problem in which the upper level optimization problem is concerned with minimizing the learner's cost post adversarial data transformation, while the lower level optimization problem of finding the optimal data transformation becomes the constraint of the upper level problem. They show a special case of the NP-hard bilevel optimization problem in which the feature space is unrestricted and the adversary's loss function is both convex and continuously differentiable with respect to the predicted label. They solve the Stackelberg games with three instances of loss functions: worst-case loss, linear loss, and logistic loss. They show that the Stackelberg model with logistic loss is more robust against adversarial attacks compared with the baseline methods.

2.3 | Modeling adversarial learning as single-leader-multifollower games

So far all the discussions on existing work are limited to the case where there is only one type of adversaries. Hence, they can be modeled as one player in the game. In more general settings, there can be a number of adversaries of various types that are unknown to the single learner. For example, in the application of malware detection, machine learning-based antivirus software may have to face hackers that are capable of various types of obfuscation attacks. Some hackers may be interested in data obfuscation with aliasing and data encoding; and others may transform structured malware programs into spaghetti code through control flow obfuscation with techniques such as aggregation/deaggregation, reordering code and data, flattening control flow graphs, and adding spurious computation.

It is hard to implement a learning technique with a single predictive model to effectively defend against every possible type of adversary. As machine learning methods become increasingly popular in the security domain, machine learning-based techniques have to be accustomed to facing the challenges from many unknown types of adversaries. A Stackelberg game in which the leader may face one of many types of followers is referred to as a *Bayesian Stackelberg game*. Finding optimal strategies in a Bayesian Stackelberg game is NP-hard (Conitzer & Sandholm, 2006). Recently, Paruchuri et al. (2008) present an efficient algorithm, referred to as Decomposed Optimal Bayesian Stackelberg Solver (DOBSS) for finding the optimal strategy for the leader in these games. They test their method in a single-leader-single-follower Bayesian Stackelberg game that models interactions between a security agent and a criminal of an uncertain number of types. The security agent as the leader must commit to its strategy and stick with the same strategy. As the follower, the criminal can play its best strategy given knowledge about the security agent's strategy. They solve the Bayesian Stackelberg game as a mixed integer linear programming problem. Solving Bayesian Stackelberg games efficiently is crucial in many security domains.

In our recent work (Zhou & Kantarcioglu, 2016), we present a single-leader-multiple-follower game between a learner and multiple adversaries of different types. The goal of the adversaries is to modify data so as to increase the learner's classification error. When modifying malicious instances, adversaries may use different strategies to corrupt the test data. Some adversaries are very aggressive in terms of modifying data, while others may attack mildly. Some adversaries may modify only malicious instances, while others may modify both malicious and benign instances. We introduce a nested Stackelberg game framework to simultaneously deal with malicious data corruption and unknown adversary types. The game framework consists of a set of lower level Stackelberg games and an upper level Bayesian Stackelberg game. The lower level single leader single follower (SLSF) Stackelberg games are solved and the solutions are used as the pure strategies for the learner. The upper level SLMF Bayesian Stackelberg game is solved to find the optimal mixed strategy for the learner. The lower level SLSF Stackelberg games take care of situations where training and test data are not necessarily i.i.d. in practice. The upper level Bayesian Stackelberg game takes care of situations where the learner has to face multiple adversaries of various types. The equilibrium strategy for the learner is an optimal mixed strategy in which each pure strategy is chosen with a probability. In other words, at each play, the learner may choose any predictive model to make the predictions, and the probability of choosing a predictive model is determined as soon as the optimal equilibrium strategies for the Bayesian Stackelberg game are found. It is quite effective for the learner to play a mixed strategy consisting of a set of predictive models with assigned probabilities. The optimal mixed strategy introduces randomness to the solution, and therefore makes it more difficult for adversaries to attack the underlying learning algorithms. In a game where there is a single leader, followed by n followers of m follower types, we assume the n followers are independent of each other and their actions have no influence on the decisions of the other players.

single leader single follower
即，一个学习系统 与 一个类型的敌手之间的博弈

With this assumption, we can reduce the problem to solving m instances of the SLSF game in which the leader is unaware of the adversary's type.

将问题简化、归约为 ：解决 m个 SLSF 博弈 的实例
其中，leader （学习系统）并不知道敌手具体的 攻击类型

3 | DISCUSSION

silver bullet :银弹
代指： 有效地解决方案，高招

The competition between the learner and the adversary often turns into an arms race. As the competition continues, one player's action is often answered with a more advanced and sophisticated response from the other player. It remains as an open problem whether a silver bullet solution exists to stop such seemingly never-ending competition, especially when the adversary's attack strategies are inexhaustible.

表面上，看似

With the existing techniques developed for robust adversarial learning, and perhaps many more to come, which technique should we adopt? The majority of the existing work focuses on modeling the problem as a game between the learner and the adversary. But as always the devil is in the details. When we apply game theory, we assume the players in the game would interact rationally. This assumption is not necessary true in real life, especially in security domains where we often have to model interactions with humans, for example, a game between airport security and terrorists. The behavior of human players can be highly unpredictable. Unpredictability does not necessarily imply irrationality or stupidity. However, it is helpful to understand that when your opponents are indeed playing poorly, staying with the equilibrium strategy will make you lose the opportunity to exploit your opponent's weaknesses. As a result, you may expect payoffs that are significantly lower than otherwise.

The same argument applies to playing mixed strategies. In the airport security example, without knowing how the terrorists would respond to its defense, it is best for the airport to deploy its security patrols randomly by playing a mixed strategy. However, when there is good reason to believe that the terrorists are not following the playbook on the equilibrium path, playing pure strategies may provide better airport security. On the other hand, it is always unwise to underestimate our opponent. It is risky to step away from the equilibrium strategy, therefore unless the odds is adequately in our favor, playing equilibrium strategies will be our best choice. With that said, it does not mean that we cannot take measured risks when we perceive a great opportunity of doing significantly better, or in cases of real emergency or human suffering.

It is worth noting that game theoretic solutions are not always feasible. In some cases, an equilibrium may not always exist; and even when it does exist, playing the equilibrium strategy may involve solving problems that are computationally intractable (Halpern & Pass, 2015). Future research can pursue games with incomplete and asymmetric information, and players occasionally not making completely rational choices.

As a related topic, game theory is used in generative models such as generative adversarial networks (GANs) (Goodfellow et al., 2014). GANs trained on benchmark image datasets are capable of generating visually realistic sample images (Chen et al., 2016). GANs are trained by searching for a Nash equilibrium of a two player noncooperative game. Research is ongoing in improving the stability and convergence of GANs (e.g., Denton, Chintala, Szlam, & Fergus, 2015; Salimans et al., 2016). Another related topic is to make deep neural networks robust against minor perturbations, often through transforming the training images (e.g., Guo, Rana, Cisse, & van der Maaten, 2018; Papernot & McDaniel, 2017; Tramèr, Kurakin, Papernot, Boneh, & McDaniel, 2018). Detailed discussion on the two related topics is beyond the scope of this survey paper.

The game theoretic studies provide important insights for building resilient learning algorithms. A learning algorithm resilient to adversarial attacks should have at least one of the following properties:

1. Robust attribute selection: When a learning algorithm is trained on a set of properly selected features that are either costly or difficult for the adversary to modify, the remaining utility of the attack instances that can evade detection will be significantly reduced. For example, Kantarcioglu et al. (2011) show that a more robust classifier can be built on an optimal (most costly) set of attributes. A classifier has much improved equilibrium performance using the most costly attributes. Being able to select the features used in a learning algorithm ultimately means the learner sets the tone of the game. Selecting features that discourage the adversary to generate malicious samples is one important direction to the adversarial learning research. It has the ability to stop or at least slow down the endless arm race between the learner and the adversary.
2. Conservative strategy: The learning algorithm can adapt to malicious samples. This can be done through periodical retraining on samples with new variants of malicious samples. This is rather a passive response from the defender's perspective and may not be desirable in certain security domains where proactive actions are always required. In situations where risks are high, conservative strategies may turn out to be a good choice. How conservative a classifier needs to be can be determined by the classifier's equilibrium performance in the game. Zhou et al. (2012) demonstrate how a SVM classifier adapts to adversarial attacks by playing a conservative strategy in anticipation of attacks as shown in Figure 2. The dashed line represents the decision boundary of a standard SVM classifier and the solid line is the decision boundary

虚线 决策边界 标准支持向量机分类器 实线

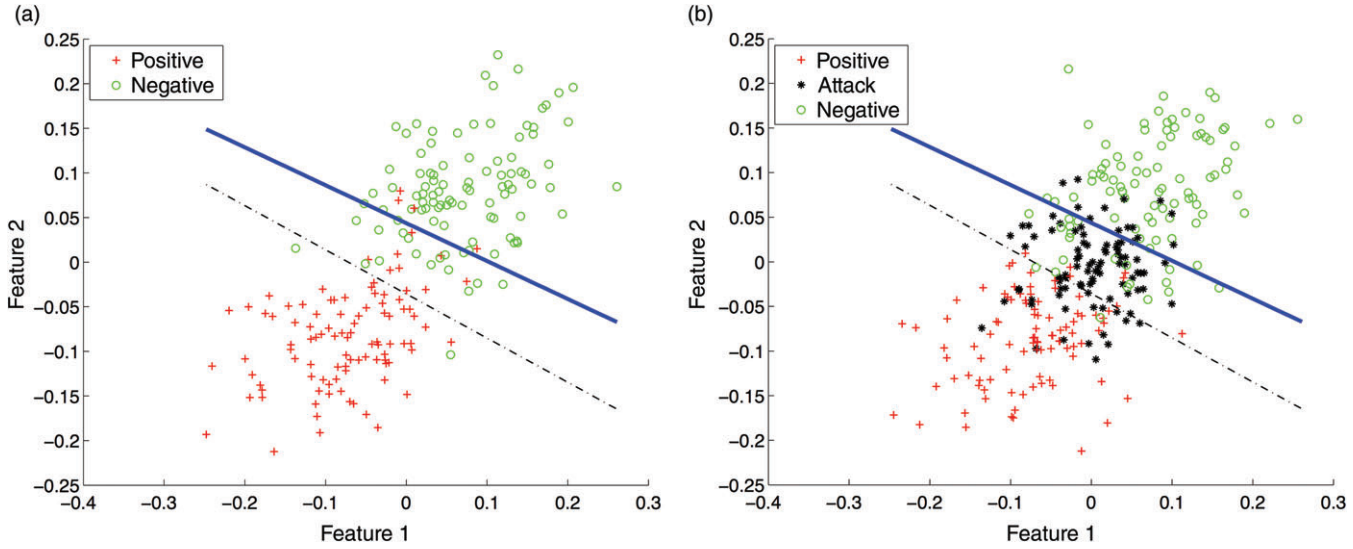

FIGURE 2 Standard support vector machine versus conservative support vector machine

保守支持向量机分类器

of a conservative SVM classifier. As can be seen, the majority of the malicious variants (shown as asterisks in the right plot) evade detection by crossing the decision boundary (dashed line) of the standard SVM, while in the meantime fail to foil a more conservative SVM classifier.

3. Randomness: It is much harder to "protect" a stationary decision boundary. A learning algorithm is more robust against adversarial attacks if it randomly updates its decision boundary. This is similar to having a mixed strategy as demonstrated by Zhou and Kantarcioglu (2016) in their nested game framework, in which each pure strategy of the computed mixed strategy is chosen with a probability to make a prediction. By introducing randomness to the decision function, the adversary does not have complete information about the learner's strategy, and the learning algorithm will make it more difficult for the adversary to generate an effective malicious variant. 恶意变体

过度拟合

In addition, a robust learning algorithm should generally avoid overfitting. Overfitting occurs when an algorithm fits the training data perfectly but fails to generalize to the unseen data. When a learning algorithm overfits, it may fail to detect a large part of unseen malicious samples. These undetected samples can be further used as the seeds to transform malicious samples currently blocked by the overfitted classifier.

省略that的定语从句，修饰 malicious samples

表目的

4 | CONCLUSIONS

Machine learning algorithms become increasingly popular in many real applications because of their impressive accuracy and potency in making decisions. However, as their popularity grows, machine learning algorithms become the victim of their own success. They are often the target of adversarial attacks. Encounter of adversaries is not limited to security domain. For example, learning algorithms for smart advertising or webpage ranking may be targeted by adversaries that are motivated to gain financial advantages. In applications that typically operate in adversarial environment, challenges from facing adversarial attacks are more apparent. Making machine learning algorithms more resilient to adversarial attacks has become an immediate necessity.

成为 当务之急

Robust learning techniques have been studied and their initial success is encouraging. Major progress has been made in the game theoretic modeling of adversarial machine learning problem. Several different models have been proposed: Some are designed to find a set of high quality features that make adversarial attacks less amenable, while some are proposed to address worst case scenarios. Some techniques assume the learner and the adversary are making simultaneous moves, while others assume the game is played sequentially with one player acting as the leader and the other the follower. Game theoretic models are also proposed to address more complex situations where there are many adversaries of different types. The advantage of game theory-based techniques include an equilibrium solution that once played by rational players may potentially end the evolutionary arms race, since the equilibrium strategies are acceptable and sufficient to both players and none of them would have the incentive to change. However, equilibrium solutions may not exist and sometimes finding them is computationally intensive. 计算密集型

Whether or not there exists a panacea for all challenges from adversaries in using machine learning techniques remains as an important open problem.

ACKNOWLEDGMENT

This work is supported by ARO grant W911NF-171-0356.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLES

Game-theoretic computing in risk analysis

REFERENCES

Barreno, M., Bartlett, P. L., Chi, F. J., Joseph, A. D., Nelson, B., Rubinstein, B. I., . . ., Tygar, J. D. (2008). Open problems in the security of learning. In *Proceedings of the 1st ACM workshop on Workshop on AISec* (pp. 19–26). New York, NY: ACM.

Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. *Machine Learning*, *81*(2), 121–148.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., . . ., Roli, F. (2013). Evasion attacks against machine learning at test time. In *Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases—Volume Part III*, ECMLPKDD'13 (pp. 387–402). Berlin, Germany: Springer-Verlag.

Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12 (pp. 1467–1474), Madison, WI: Omnipress.

Bruckner, M., & Scheffer, T. (2009). Nash equilibria of static prediction games. In *Advances in Neural Information Processing Systems*, Red Hook, NY: Curran Associates Inc.

Bruckner, M., & Scheffer, T. (2011). Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2172–2180), Red Hook, NY: Curran Associates, Inc.

Conitzer, V., & Sandholm, T. (2006). Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, EC '06 (pp. 82–90). New York, NY: ACM.

Dalvi, N., Domingos, P., Mausam, S., and Verma, D. (2004). Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04 (pp. 99–108). New York, NY: ACM.

Dekel, O., & Shamir, O. (2008). Learning to classify with missing and corrupted features. In *Proceedings of the International Conference on Machine Learning* (pp. 216–223). New York, NY: ACM.

Dekel, O., Shamir, O., & Xiao, L. (2010). Learning to classify with missing and corrupted features. *Machine Learning*, *81*(2), 149–178.

Denton, E. L., Chintala, S., Szlam, A., & Fergus, R. (2015). Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems* (pp. 1486–1494), Cambridge, MA: MIT Press.

El Ghaoui, L., Lanckriet, G. R. G., & Natsoulis, G. (2003). *Robust classification with interval data.* Technical Report UCB/CSD-03-1279, EECS Department, University of California, Berkeley.

Globerson, A., & Roweis, S. (2006). Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06 (pp. 353–360). New York, NY: ACM.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . ., Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680), Red Hook, NY: Curran Associates, Inc.

Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018). Countering adversarial images using input transformations. To appear in Proceedings of the Sixth International Conference on Learning Representations.

Halpern, J. Y., & Pass, R. (2015). Algorithmic rationality: Game theory with costly computation. *Journal of Economic Theory*, *156*, 246–268.

Kantarcioglu, M., Xi, B., & Clifton, C. (2011). Classifier evaluation and attribute selection against active adversaries. *Data Mining and Knowledge Discovery*, *22*, 291–335.

Koscher, K., Czeskis, A., Roesner, F., Patel, S., Kohno, T., Checkoway, S., . . ., Savage, S. (2010). Experimental security analysis of a modern automobile. In *Proceedings of the 2010 I.E. Symposium on Security and Privacy*, SP '10 (pp. 447–462). Washington, DC: IEEE Computer Society.

Lanckriet, G. R. G., Ghaoui, L. E., Bhattacharyya, C., & Jordan, M. I. (2002). A robust minimax approach to classification. *Journal of Machine Learning Research*, *3*, 555–582.

Liu, W., & Chawla, S. (2010). Mining adversarial patterns via regularized loss minimization. *Machine Learning*, *81*, 69–83.

Lowd, D., & Meek, C. (2005). Adversarial learning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05 (pp. 641–647). New York, NY: ACM.

Papernot, N., & McDaniel, P. (2017). Extending defensive distillation. Poster session presented at the 38th IEEE Symposium on Security and Privacy.

Paruchuri, P., Pearce, J. P., Marecki, J., Tambe, M., Ordonez, F., & Kraus, S. (2008). Playing games for security: An efficient exact algorithm for solving Bayesian Stackelberg games. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '08 (pp. 895–902), Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *Advances in Neural Information Processing Systems* (pp. 2234–2242), Red Hook, NY: Curran Associates Inc.

Scholkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT Press.

Teo, C. H., Globerson, A., Roweis, S. T., & Smola, A. J. (2007). Convex learning with invariances. In *Advances in Neural Information Processing Systems*, Red Hook, NY: Curran Associates Inc.

Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. To appear in Proceedings of the Sixth International Conference on Learning Representations.

Wagner, D., & Soto, P. (2002). Mimicry attacks on host-based intrusion detection systems. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, CCS '02. (pp. 255–264). New York, NY: ACM.

Xu, W., Qi, Y., & Evans, D. (2016). Automatically evading classifiers: A case study on PDF malware classifiers. In L. Bauer & K. O'Donoghue (Eds.), *The network and distributed system security symposium 2016*. San Diego, CA: Internet Society.

Zhou, Y., & Kantarcioglu, M. (2016). Modeling adversarial learning as nested Stackelberg games. In J. Bailey, L. Khan, T. Washio, G. Dobbie, J. Z. Huang, & R. Wang (Eds.), *Advances in Knowledge Discovery and Data Mining - 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19–22, 2016, Proceedings, Part II, volume 9652 of Lecture Notes in Computer Science* (pp. 350–362). New York, NY: Springer.

Zhou, Y., Kantarcioglu, M., Thuraisingham, B., & Xi, B. (2012). Adversarial support vector machine learning. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12. (pp. 1059–1067). New York, NY: ACM.

How to cite this article: Zhou Y, Kantarcioglu M, Xi B. A survey of game theoretic approach for adversarial machine learning. *WIREs Data Mining Knowl Discov*. 2018;e1259. https://doi.org/10.1002/widm.1259