Computing Reviews
Association for Computing Machinery
ThinkLoud

TODAY'S ISSUE | HOT TOPICS | SEARCH | BROWSE | RECOMMENDED | MY ACCOUNT | LOGIN
ALL

Hot Topics | essay

Search

Adversarial Learning: Mind Games with the Opponent

Murat Kantarcioglu
University of Texas at Dallas
Bowei Xi
Purdue University
Yan Zhou
University of Texas at Dallas

1. Introduction

Machine learning algorithms have created the sensation of working with experts in an endless list of fields—advertisers, detectives, doctors, chess masters, soccer players, loan creditors, salespeople, entertainment critics, drivers, and even matchmakers. Machines with intelligence have dazzled the human mind with what they can do with data, and how much better they are at it. These smart algorithms can learn from our observations about the world and accomplish tasks that would appear to be insurmountable hurdles to humans. However, machine learning algorithms may be terrible in situations when the conditions are not right for them, for example, when they face an adversary.

In many applications, especially in cyber security, machine learning-based tools have to face challenges from their natural opponents. For example, hackers may obfuscate their malicious code to evade detection by machine learning-based anti-virus software, or spammers often disguise their spam messages to defeat spam filters by inserting excerpts from daily newspapers. The following two studies typify the kind of challenge machine learning algorithms face in real-world applications.

- **PDFrate** is a machine learning-based PDF malware classifier available online. The classifier is trained on 5,000 malicious PDF documents and over 100,000 benign documents using random forest--an ensemble method that builds each of its individual trees on a random subset of features. Despite the claim that the system has above 99 percent accuracy with a less than 0.2 percent false positive rate [1], Šrndić and Laskov [2] successfully attack the classifier by injecting dummy content into malicious PDF documents. They manage to reduce the classification score from nearly 100 percent to approximately 33 percent. More importantly, their work reveals that potential defense mechanisms are only effective if the attack exactly matches the anticipated ones.

- **Facial recognition** is another typical real-world application of machine learning algorithms that may face adversarial attacks. Facial recognition technologies have broad applications in access control, biometric identification, and surveillance. Their accountability has a great impact on making security decisions in uncertain situations. Despite assurance of high accuracy on normal facial images, Sharif et al. [3] demonstrate how fragile a machine learning-based facial recognition system can be when facing adversarial attacks. By adding a simple eyeglass frame, actress Reese Witherspoon is misclassified as actor Russell Crowe.

2. Vulnerabilities of Machine Learning Algorithms

The strategy of the adversaries is to force these machine experts to step out of their comfort zones. Here's how. Machine learning algorithms are trained on a set of collected data and are expected to perform well when competing in an arena with the same conditions. However, the adversaries can turn the tables by modifying the data, either at training time (when models are built) or test time (when machine learning models are deployed). In either case, machine learning algorithms are trained on data of one distribution, but tested on a different data distribution. When this happens, the performance of machine learning algorithms starts to derail. Figure 1 illustrates an example of how things can go wrong when an adversary strikes by modifying the input to a support vector machine (SVM)--a machine learning algorithm that outputs a hyperplane that separates two classes of data with a maximum margin.

Related Resources:

Books

Adversarial reasoning: computational approaches to reading the opponent's mind Kott, A; McEneaney, W. M. (Eds.), 2007

Conferences and Workshops

ACM Workshop on Artificial Intelligence and Security (AISec): annual workshop for security, privacy, AI and machine learning researchers. AISec 2016 especially focused on learning in game-theoretic adversarial environments, among other topics.

Neural Information Processing Systems (NIPS) 2016 Workshop on Adversarial Learning: workshop focused on adversarial training, in which "a set of machines learn together by pursuing competing goals."

Data Science for Cyber-Security Conference (DSCB): upcoming conference (September 2017) with a focus on cutting-edge research on data science in cyber-security applications, including machine learning, big data analytics for network modeling, and forensics, as well as other similar topics.

Videos

Adversarial data mining: big data meets cyber security – Part 1
Kantarcioglu, M.; Xi, B. *ACM CCS 2016*

Reviews

Game theory with engineering applications
Bauso D., SIAM, 2016.

Adversarial reasoning
Kott A., McEneaney W., Chapman & Hall/CRC, 2006.

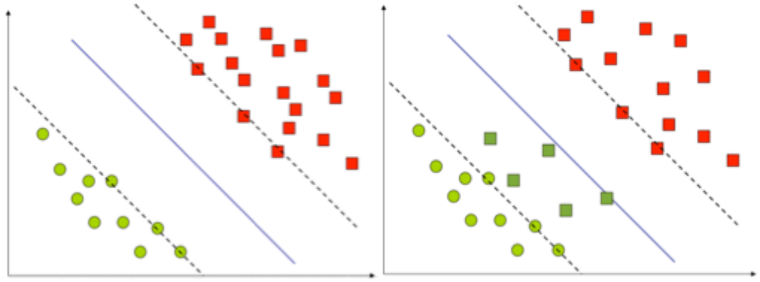

Figure 1: Adversarial attacks against a support vector machine on a two-dimensional dataset. The red squares represent malicious samples, and the green dots are benign samples. The green squares are malicious samples modified by the adversary. The blue line is the decision boundary of the support vector machine.

In Figure 1, on the left we show a normal machine learning task where the goal is to classify samples into two classes. On the right, we show how modified malicious samples (marked as green squares) pass screening by the SVM classifier.

All standard learning algorithms are vulnerable to adversarial attacks because they are all built on the i.i.d. assumption, that is, training and test data are independently and identically distributed. Uncertainty created by adversaries in the input to machine learning algorithms adds doubt to the “knowledge” they have learned. In the presence of adversaries, the optimal solutions a learning algorithm produces become vulnerable and may be harnessed by the adversaries to defeat the learning algorithm. Adversaries can “poison” a small set of training samples to mislead the learning algorithm if they have the ability to access the training data. For example, Battista Biggio et al. [4] studied poisoning attacks against SVMs. They inject a specially crafted training sample into the training set to maximally increase the loss incurred on a separate validation dataset. In reality, for a learning system to voluntarily accept a poisoned training sample, a chain of scenarios would have to unfold: adversaries gaining privileges to access the training set, attacks going unnoticed when tested against the clean validation dataset, and adversaries having the ability to control the labeling of the poisoned sample. Without these necessary conditions that allow for modifying the training data, many adversaries are more interested in disguising their malicious samples at test time to evade detection by the learning system. Adversarial attacks are most commonly encountered in security sensitive and economically driven domains. This class of learning problems where resilience to adversarial attacks is critical is known as adversarial learning.

Deep learning, a popular topic in the recent development of machine learning, is not spared from poor performance when it faces adversarial attacks. For example, Christian Szegedy et al. [5] show that deep neural networks are prone to slight perturbations on the input data. In their experiment, deep neural networks failed to correctly classify mildly perturbed images that humans have no difficulty classifying accurately. Ian Goodfellow et al. [6] present a perturbation technique that modifies all input dimensions of a sample by a small quantity in the direction of the fast gradient sign they computed. Nicolas Papernot et al. [7] also present algorithms for crafting adversarial samples. They show that their algorithms are able to reliably compute adversarial samples that can fool a deep neural network with a 97 percent success rate. More important, they only modify 4.02 percent of the input per sample on average. It is worth noting that all of the attacks on deep neural networks discussed here assume a white-box scenario in which the adversary knows everything about the trained deep neural networks including the structure and all of the parameters. Mahmood Sharif et al. [3] have demonstrated initial success under the black-box scenario where the adversary can only query trained deep neural networks to get scores for the candidate samples. Their algorithm iteratively searches for the qualified candidate by applying particle swarm optimization, a heuristic and stochastic optimization algorithm that mimics the behavior of a swarm of birds.

3. Adversarial Learning

To build a robust learning system for adversarial situations, we have to take into account not only the rigor of learning algorithms in the presence of malicious attacks, but also the mathematical model of adversarial conflict. We define adversarial learning as follows:

Adversarial learning is the study of robust machine learning algorithms developed to effectively counter various adversarial attacks.

A taxonomy of adversarial attacks against machine learning algorithms has been created according to the attacker’s capability, the type of security violation, and the attacker’s intent [8]:

- *Causative vs. exploratory attacks*: causative attacks misguide learning through training data, while exploratory attacks exploit classification-time errors without affecting training.
- *Integrity vs. availability attacks*: integrity attacks cause misclassification of malicious instances (false negatives), and availability attacks sabotage the reliability of learning by causing misclassification of benign instances (false positives).
- *Targeted vs. indiscriminate attacks*: targeted attacks focus on misclassification of a specific instance, while indiscriminate attacks cause reduced predictive accuracy on a pool of instances.

In general, adversarial learning algorithms robust to adversarial attacks should be able to recognize various deceptions, think as the enemy by modeling adversarial conflict, and predict the adversary's intent.

4. Countermeasures Against Adversarial Attacks

Adversarial learning can be naturally modeled as a two-player game between a learning system and an adversary. When one player's gain is the other player's loss, they are said to play a zero-sum game. A zero-sum game is a strictly competitive game in which two players have completely opposing preferences. A simple example of a typical zero-sum game is the well-known children's game Rock-Paper-Scissors.

In earlier stages of research, adversarial learning is often modeled as a zero-sum game. In the game, the learner needs to come up with a strategy to achieve the greatest average payoff by carrying out a worst-case analysis, in which the learner assumes the adversary knows his strategy and will play optimally against it. Then, the learner's worst-case loss is minimized over all possible data points in conformity with predefined constraints. For example, Amir Globerson and Sam Roweis [9] solve an optimal SVM learning problem to deal with malicious feature deletion on test data. They search for the zero-sum minimax strategy that minimizes the hinge loss of the SVM. Laurent El Ghaoui et al. [10] present a minimax strategy for a similar problem in which training data is bounded by hyper-rectangles. Our earlier work also fits in this line of research [11]. We present a minimax strategy for input data constrained by two attack models. The attack models are defined in terms of the adversary's capabilities of modifying data. Our solutions minimize the worst-case loss corresponding to the two attack models. Michael Brückner and Tobias Scheffer [12] also studied static prediction games in which they assume both players commit to their strategies simultaneously and search for a unique Nash equilibrium solution. Cost-sensitive opponents have also been studied by Nilesh Dalvi and his colleagues [13] in a game between a classifier and an adversary. Given a cost function that estimates the cost of transforming an instance, the adversary transforms an instance for which the cost is minimized.

In many real applications, the adversarial learning problem is more appropriately modeled as a sequential game between two players. First, one player takes initiatives to maximize the threat to its opponent, and then the opponent answers by deploying the most effective countermeasures. One player must commit to its strategy before the other player responds. The advantage the responding player has is partial or complete information of the first player. The responding player can therefore play its optimal strategy against its opponent. This type of game is known as a Stackelberg game in which the first player is the leader and its opponent is the follower.

Adversarial learning research in this area falls into two categories, depending on who plays the role of the leader: the learner or the adversary. In our earlier work [14], we solve for a Stackelberg equilibrium using simulated annealing to discover an optimal set of attributes to build machine learning models. Similar work has also been done by Wei Liu and Sanjay Chawla [15]. The difference between their research is that the former assumes both players know each other's payoff function, while the latter relaxed the assumption and only the adversary's payoff function is required. In both cases, the adversary is the leader whose strategies are stochastically sampled while the learner is the follower that searches for an equilibrium given its knowledge about the adversary.

In other settings, the learner is more likely to commit to a strategy before the adversary takes its actions. The adversary's response is optimal given that it has some knowledge about the classifier's strategy. For example, email service providers usually have spam filters installed on the server side before providing services to the end user. Sophisticated spammers would obtain firsthand knowledge about the statistical tools implemented in the spam filters before sending out massive amounts of spam. This can be done by probing spam filters with a plentiful supply of carefully crafted email messages. These messages are specifically designed for detecting the decision boundaries of the spam filters. Example solutions to this type of Stackelberg game are presented by Michael Brückner and Tobias Scheffer [16]. They define a Stackelberg prediction game in which the learner minimizes its loss knowing that the adversary is playing its optimal strategy. The adversary's strategy is optimal if it is among the solutions that minimize the adversary's loss given the learner's strategy.

In more realistic settings, there are possibly many adversaries of various unknown types. For example, in email spam filtering, some spammers are interested in the successful delivery of spam contents to the end user, while others attempt to flood the network by performing denial-of-service attacks; some spammers try to corrupt both training and test data at high costs, while others choose to alter only test data at a much lower cost; some spammers can modify both spam and legitimate email, while others have fewer privileges and are not entitled to access legitimate email. It is hard to implement a spam filter with a single predictive model to effectively counter every possible type of adversary. Therefore, a single leader in Stackelberg games may have to face many adversaries of different types. Praveen Paruchuri and his colleagues [17] present a single-leader-single-follower (SLSF) Bayesian Stackelberg game to model interactions between a security agent and a criminal of an uncertain number of types. The security agent has only one type and must commit to its strategy first and stick with it. The criminal plays its best strategy given knowledge about the security agent's strategy. They solve the Stackelberg game as a mixed integer linear programming problem. In our recent work [18], we study a single-leader-multiple-followers (SLMF) game between a learner and multiple adversaries. We consider a malicious data modification problem where adversaries may use different strategies to corrupt the test data. We present a nested Stackelberg game framework to handle both data corruption and unknown types of adversaries. The game framework consists of a set of SLSF Stackelberg games and an SLMF Bayesian Stackelberg game. The low-level SLSF Stackelberg game takes into consideration that training and test data are not necessarily identically distributed in practice. The top-level Bayesian Stackelberg game consists of one learner and multiple adversaries of various types. When there are adversaries of multiple types, instead of settling on one learning model by playing a pure strategy, it is more practical for the learner to play a mixed strategy consisting of a set of learning models with assigned probabilities. The optimal solution to the Bayesian Stackelberg game introduces randomness to the solution, and hence increases the difficulty of attacking the underlying learning algorithms.

5. Concluding Remarks

An important note is that adversarial learning often becomes an arms race between the learner and the adversary as the competition continues. One player's improvement is often followed by a more sophisticated countermeasure from the other player. Solving the problem would lead to constant model improvement or update given that the adversary's attack strategies are inexhaustible. Initial results indicate that building machine learning models using useful attributes that are hard to modify for the attacker is a good starting point. Still, it remains as an open problem whether a silver-bullet solution exists to stop such seemingly never-ending competition.

SHARE

Created: Aug 7 2017
Last updated: Aug 7 2017

References

1) Smutz, C.; Stavrou, A. Malicious PDF detection using metadata and structural features. In *Proc. of ACSAC '12* (2012), 239–248.

2) Šrndić, N.; Laskov, P. Practical evasion of a learning-based classifier: a case study. In *Proc. of SP '14* (2014), 197-211.

3) Sharif, M.; Bhagavatula, S.; Bauer, L.; Reiter, M. K. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In *Proc. of CCS '16* (2016), 1528-1540.

4) Biggio, B.; Nelson, B.; Laskov, P. Poisoning attacks against support vector machines. In *Proc. of ICML '12* (2012), 1807–1814.

5) Szegedy, C.; Zaremba, W.; Sutskever, I; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. In *Proc. of ICLR '14* (2014), 10 pages.

6) Goodfellow, I. J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. In *Proc. of ICLR '15* (2015), 11 pages.

7) Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; Swami, A. The limitations of deep learning in adversarial settings. In *Proc. of SP '16* (2016), 372-387.

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

9) Globerson, A.; Roweis, S. Nightmare at test time: robust learning by feature deletion. In *Proc. of ICML '06* (2006), 353–360.

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

11) Zhou, Y.; Kantarcioglu, M.; Thuraisingham, B.; Xi, B. Adversarial support vector machine learning. In *Proc. of KDD '12* (2012), 1059–1067.

12) Brückner, M.; Scheffer, T. Nash equilibria of static prediction games. In *Proc. of NIPS '09* (2009), 171-179.

13) Dalvi, N.; Domingos, P.; Mausam; Sanghai, S.; Verma, D. Adversarial classification. In *Proc. of KDD '04* (2004), 99–108.

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

15) Liu, W.; Chawla, S. A game theoretical model for adversarial learning. In *Proc. of ICDMW '09* (2009), 25–30.

16) Brückner, M.; Scheffer, T. Stackelberg games for adversarial prediction problems. In *Proc. of KDD '11* (2011), 547–555.

17) Paruchuri, P.; Pearce, J. P.; Marecki, J.; Tambe, M.; Ordonez, F.; Kraus, S. Playing games for security: an efficient exact algorithm for solving Bayesian Stackelberg games. In *Proc. of AAMAS '08* (2008), 895-902.

18) Zhou, Y.; Kantarcioglu, M. Modeling adversarial learning as nested Stackelberg games. In *Proc. of PAKDD '16* (2016), 350-362.

REVIEWER'S AREA | MASTHEAD | SUBSCRIBE | NEWS | TIPS | HELP | CONTACT US

Select Language ▼ Powered by Google Translate