Deceptive and Counter-Deceptive Machines
Papers from the AAAI 2015 Fall Symposium

Toward Adversarial Online Learning and the Science of Deceptive Machines

Myriam Abramson
Naval Research Laboratory, Code 5584
Washington, DC 20375
myriam.abramson@nrl.navy.mil

Abstract

Intelligent systems rely on pattern recognition and signature-based approaches for a wide range of sensors enhancing situational awareness. For example, autonomous systems depend on environmental sensors to perform their tasks and secure systems depend on anomaly detection methods. The availability of large amount of data requires the processing of data in a "streaming" fashion with online algorithms. Yet, just as online learning can enhance adaptability to a non-stationary environment, it introduces vulnerabilities that can be manipulated by adversaries to achieve their goals while evading detection. Although human intelligence might have evolved from social interactions, machine intelligence has evolved as a human intelligence artifact and been kept isolated to avoid ethical dilemmas. As our adversaries become sophisticated, it might be time to revisit this question and examine how we can combine online learning and reasoning leading to the science of deceptive and counter-deceptive machines.

1 Introduction

Intelligent systems rely on pattern recognition and signature-based approaches for a wide range of sensors enhancing situational awareness. For example, autonomous systems depend on environmental sensors to perform their tasks and secure systems depend on anomaly detection methods or behavior identification methods. In addition, the voluminous amount of digital traces enables the construction of "cognitive fingerprints" (Guidorizzi 2013; Abramson 2015; Robinson 2010) for multi-factor authentication with personalized user profiles characterizing our unique interaction with technology. The availability of large amount of data requires the processing of data in a "streaming" fashion with online algorithms. Yet, just as online learning can enhance adaptability to a non-stationary environment, it introduces vulnerabilities that can be manipulated by adversaries to achieve their goals while evading detection. The arms race between spam and spam filters is a precursor to something more insidious that will permeate the fabric of our technology-mediated interactions in cyberspace. In cyber-security, we use machine learning to anticipate surprises, such as zero-day exploits, but our adversaries learn themselves to avoid detection creating more vulnerabilities.

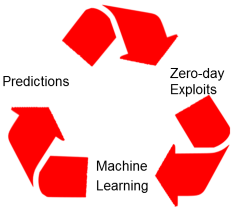


Figure 1: Arms race in cyber-security

As a result, the human-in-the-loop modifies the algorithms, usually by adding/removing features, or restart the learning process with assumed non-tainted examples and the cycle continues (Fig.1). The cognitive biases of the human-in-the-loop are a source ripe for deception by our attackers (Marble et al. 2015). In addition, increased automation, due to increased and faster connectivity, raises the possibility that machines will deceive each other. Although human intelligence might have evolved from social interactions (Dunbar 2003), machine intelligence has evolved as a human intelligence artifact, inheriting cognitive biases through training, and has been kept isolated to avoid ethical dilemmas. The claim of this paper is that, as our adversaries become increasingly sophisticated, it might be time to revisit this question and examine the consequences, limitations, and implications of combining online learning and reasoning leading to the science of deceptive and counter-deceptive machines. Will this result merely in robust machine intelligence or in a new, more fluid, kind of machine intelligence?

This position paper on proposed future research is organized as follows. Section 2 introduces adversarial learning and its relation to deceptive machines. Section 3 introduces the computational formalism for a counter-measure approach that includes deceptive and counter-deceptive actions. Section 4 presents related work in "secure" machine learning addressing machine learning vulnerabilities in adversarial conditions. Finally, we conclude in Section 5 with a summary and a discussion on the limitations of deceptive learning machines.

2 Adversarial Learning

Adversarial machine learning is a game against an adversarial opponent (Huang et al. 2011; Lowd and Meek 2005) who

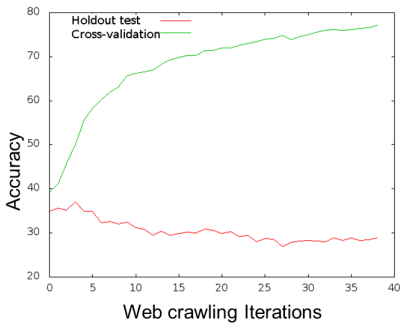


Figure 2: Self-deception with semi-supervised learning and self-teaching

Figure 3: Adversarial drift over time in a user authentication experiment where impostor data that evaded detection are slowly ingested by the algorithm.

tries to deceive the algorithm into making the wrong prediction by manipulating the data. This deception occurs in two ways:

- Temporal drift: behavior changes over time requiring retraining of the model. Adversaries can take advantage of this adaptability by injecting poisonous examples masquerading as real (camouflage). Since there is no clear separation between training and testing in online learning algorithms, rather testing become training (given bandit feedback), an adversarial scenario occurs where the next label in the sequence is different than the one predicted (Mohri, Rostamizadeh, and Talwalkar 2012). This has led to the development of mistake-bound online algorithms. Consequently, an aspect of adversarial learning is how to exploit the cognitive biases of the human-in-the-loop or the algorithm to mislabel examples. The danger with bandit feedback is self-deception, that is the reinforcement of minor errors due to ambiguity. Figure 2 illustrates the performance of semi-supervised learning using self-teaching in the genre classification of URLs iteratively provided by a Web crawler and where the labels were ambiguous in the starting training set.
- Adversarial drift: signature-based approaches do not distinguish between uncommon patterns and noise. Adversaries can take advantage of this inherent blind spot to avoid detection (mimicry). Adversarial label noise is the intentional switching of classification labels leading to deterministic noise, error that the model cannot capture due to its generalization bias. An experiment in user authentication from Web browsing behavior (Abramson and Aha 2013) injected a few (1%) "malicious" examples, that is examples of behavior from other users that evaded detection, in a user training set over 20 iterations (Figure 3). Cross-validation over time does not show any anomalies. In contrast, the performance quickly deteriorates in a hold-out dataset leading to the denial of authentication service locking the user out.

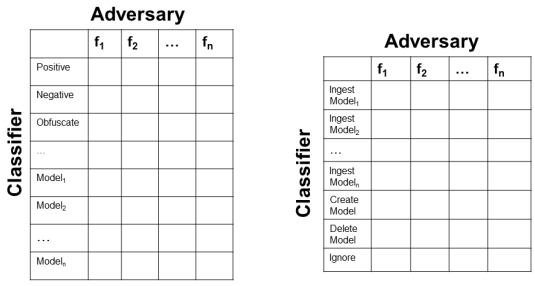
One aspect of our proposed approach is based on a game-theoretical computational framework modeling two ways of interacting with a learner in security games as follows.

- Machine probing: (1) how to find blind spots in a learner in order to manipulate predictions; and (2) how to probe a learner to divulge information about its predictability for evasion purposes. This type of interaction corresponds to exploratory attack vectors (Huang et al. 2011) which seek to gain information about a learner (e.g., its bias, window-size, features). Machine probing can occur with a non-adaptive learner but is often a precursor to machine teaching.
- Machine teaching: the main issue here is how to poison a learner to make inaccurate predictions in as few attempts as possible. This type of interaction corresponds to causative attack vectors (Huang et al. 2011) directly influencing a learner through its training data. Machine teaching is considered an inverse problem to machine learning by mapping a target model to a set of examples and has broad applications. (Mei and Zhu 2015; Zhu 2015)

The analysis of security games consists of proving (1) the convergence toward a Nash equilibrium with desired payoffs and (2) the sample complexity (i.e., the number of examples) necessary to spoof a learner online. Proving convergence toward a Nash equilibrium is essential here to guarantee that a vulnerability exists in the case of an oblivious classifier. An exponential number of resources will be expensive for an adversary (in terms of computational cost) and noticeable providing the online learning algorithms with a "natural" protection. Figure 4 illustrates the possible actions of a classifier and data manipulation of features by an adversary in a game-theoretical framework for machine probing and teaching where the payoffs are the risks of misclassification and costs of feature evaluation for the defender/classifier vs. the costs of modifying the data for the attacker. The dimensions of the data can be either the features themselves, feature subsets, or parameters of the data distribution. Analysis assumptions include (coarse) knowledge of the data and the learning algorithm.

3 Adversarial Meta-Learning

While the analysis of machine learning vulnerabilities can provide insights into future exploits, security games are games of imperfect information where the adversary adapts



(a) Machine probing (b) Machine Teaching

Figure 4: Security games with possible actions of classifier and adversarial data manipulation of features f_i.

its game to find out the payoff matrix from a limited number of interaction to "get inside" the decision cycle of the algorithm. For example, spam evolved into personalized spam or "spear phishing" messages that avoid detection by learning our preferences and the biases of a spam filter (e.g., proportion of "good words"). This evolution of malware motivates an interactive game between the classifier and the adversary to develop an adversary-aware classifier as a countermeasure.

Toward this end, we will develop a meta-learner as a wrapper to a learner that will weight its actions against an adaptive adversary evolving its tactics in response to the learner's predictions. The algorithmic framework for non-cooperative adaptive agents is based on learning in repeated games. The challenge is in learning with infinite repeated games where the number of actions for each agent is potentially infinite and with imperfect information. We will combine Monte-Carlo tree search (MCTS) with reinforcement learning because the manipulation of the exploration/exploitation tradeoff (Kocsis and Szepesvári 2006) is particularly relevant in security games. MCTS combines bandit algorithms with opponent modeling to find the game-theoretic value of a move.

Deception is one of the factors affecting rationality in games. It has been shown that adversaries do not always play the most optimal move (due to limited observations and the possibility of being detected) and defending only against optimal strategies of a non-evolving adversary (as in a minimax strategy) might be a weakness (Pita et al. 2012). In addition, in "strategic teaching" (Camerer, Ho, and Chong 2001) an adversary might not purposely play an optimal move in order to induce an adaptive learner to make the wrong prediction triggering a best response from the adversary. Consequently, starting with imprecise information, MCTS will "unroll" the relevant actions of the adversary and the classifier in simulated sequential games and backup the payoffs obtained at the leaves. Deception techniques include concealing the algorithm bias by obfuscating its predictions, exploiting the algorithm's weaknesses to engage the adversary, and forcing the adversary to reveal its intentions. Regret minimization techniques to evaluate the payoff have been proven to converge to an approximate Nash equilibrium in zero-sum games (Blum and Monsour 2007). We will define a notion of regret for secure online learning with respect to a static learner. The value of an action a, $V(a)$, is then computed as the expected sum of discounted payoffs for each successor nodes x_i, $V(a) = E(\sum_t \gamma^t V(x_t))$. The discount parameter γ, $0 < \gamma < 1$, represents the confidence of a player where, for example, an overconfident player will discount look-aheads with a low γ. MCTS has been used successfully in solving Markov decision processes and has been coupled in a multi-armed bandit setting to sample actions selectively to reduce the search space (Kocsis and Szepesvári 2006) but not in an adversarial multi-armed bandit setting (where payoffs are rigged by an adversary). MCTS will provide the framework to develop an adversary-aware meta-classifier as a counter-measure in an online setting. We will base our evaluation on the difference in performance with a non-adversary-aware learner.

4 Related Work

There has been a lot of work dealing with the co-evolution of spam and spamfilters addressing mainly Naive Bayes as the classifier. A complete information, one-shot (1-ply) game-theoretical approach augmenting a Naive Bayes spamfilter has been proposed to detect near-optimal evasion (i.e., near-optimal malicious instance in terms of cost)(Dalvi et al. 2004). A comparison between active and passive "good word" attacks in the generation of spam has shown that active attacks where an adversary evaluates the predictions of a classifier can produce optimal results (minimum number of words to change to evade detection) (Lowd and Meek 2005).

In (Brückner, Kanzow, and Scheffer 2012), conditions for a unique Nash equilibrium in repeated play between a learner and its adversary in the form of a data generator were identified in the context of spam filtering. As a result, spam could be generated to evade the spam detector assuming complete knowledge of the learner. The transformation costs of the input features were added as a regularizer in the objective function of the data generator thereby constraining spam to the training distribution.

More recently, there has been a similar arms race between malware and malware detection. Because malware slowly evolves from the repackaging of other malware and legitimate software (Zhou and Jiang 2012), it was argued that an ensemble of classifiers, one for each family of malware, was more accurate in dealing with adversarial drift (Kantchelian et al. 2013). In fact, ensemble methods such as bagging and random subspace, have been shown to be effective against poisoning attacks by reducing the influence of outliers in the training data assuming that the adversary controls only a small percentage of the data (Biggio et al. 2011). The random subspace method where meta-features are constructed from subsets of the original features enhances "feature evenness" (Kołcz and Teo 2009) where the importance of the features is hard to detect and protect against an adversary (Biggio, Fumera, and Roli 2010). The assumption is that meta-features, such as constructed by the random subspace method, non-linear kernels, or hidden nodes in a neural network, are computationally expensive to reverse engineer. However, those meta-features are constructed from

primitive features and thus can be indirectly affected.

While progress has been made in understanding the strengths and weaknesses of certain machine learning algorithms, no unified approach has been proposed for secure machine learning.

5 Conclusion

The modeling of security games in limited resource allocation problems between an attacker and defender has been very successful in balancing surveillance and control and has been deployed to protect national infrastructure (Tambe 2011). We claim that we can also model security games in machine learning in terms of detection cost to an adversary and misclassification costs to a learner. We note that deception in machines, just like deception in humans, involves presentation style and presentation of facts as the adversary's game is to predict the decision of the classifier under attack and use that ability to construct new data. Our technical approach is at the intersection of machine learning and computational game theory informed by cognitive and behavioral science. The research involves the analysis and development of security games for machine probing where an adversary seeks to evade or learn information about a classifier and machine teaching where an adversary seeks to actively modify an online learning algorithm. We also propose a countermeasure with an adversary-aware meta-learner using MCTS combining counter-factual regret minimization and randomization and including boosting and information hiding as the possible actions of a learner.

While this approach can potentially reduce the number of false alarms, an important limitation will be to re-evaluate the role of the human-in-the-loop who now uses machine learning as a passive tool. Another limitation is the added complexity of deception and its unintended consequences.

References

Abramson, M., and Aha, D. W. 2013. User authentication from web browsing behavior. In *Florida Artificial Intelligence Society FLAIRS-26*.

Abramson, M. 2015. Cognitive fingerprints. In *2015 AAAI Spring Symposium Series*.

Biggio, B.; Corona, I.; Fumera, G.; Giacinto, G.; and Roli, F. 2011. Bagging classifiers for fighting poisoning attacks in adversarial classification tasks. In *Multiple Classifier Systems*. Springer. 350–359.

Biggio, B.; Fumera, G.; and Roli, F. 2010. Multiple classifier systems under attack. In *Multiple Classifier Systems*. Springer. 74–83.

Blum, A., and Monsour, Y. 2007. Learning, regret minimization, and equilibria.

Brückner, M.; Kanzow, C.; and Scheffer, T. 2012. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research* 13(1):2617–2654.

Camerer, C.; Ho, T.; and Chong, K. 2001. Behavioral game theory: Thinking, learning and teaching.

Dalvi, N.; Domingos, P.; Sanghai, S.; Verma, D.; et al. 2004. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 99–108. ACM.

Dunbar, R. I. 2003. The social brain: mind, language, and society in evolutionary perspective. *Annual Review of Anthropology* 163–181.

Guidorizzi, R. P. 2013. Security: Active authentication. *IT Professional* 15(4):4–7.

Huang, L.; Joseph, A. D.; Nelson, B.; Rubinstein, B. I.; and Tygar, J. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 43–58. ACM.

Kantchelian, A.; Afroz, S.; Huang, L.; Islam, A. C.; Miller, B.; Tschantz, M. C.; Greenstadt, R.; Joseph, A. D.; and Tygar, J. 2013. Approaches to adversarial drift. In *Proceedings of the 2013 ACM workshop on Artificial intelligence and security*, 99–110. ACM.

Kocsis, L., and Szepesvári, C. 2006. Bandit-based monte-carlo planning. In *Machine Learning: ECML 2006*. Springer. 282–293.

Kołcz, A., and Teo, C. H. 2009. Feature weighting for improved classifier robustness. In *CEAS 09: sixth conference on email and anti-spam*.

Lowd, D., and Meek, C. 2005. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 641–647. ACM.

Marble, J. L.; Lawless, W.; Mittu, R.; Coyne, J.; Abramson, M.; and Sibley, C. 2015. The human factor in cybersecurity: Robust & intelligent defense. In *Cyber Warfare*. Springer. 173–206.

Mei, S., and Zhu, X. 2015. Using machine teaching to identify optimal training-set attacks on machine learners. AAAI.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of machine learning*. MIT press.

Pita, J.; John, R.; Maheswaran, R.; Tambe, M.; Yang, R.; and Kraus, S. 2012. A robust approach to addressing human adversaries in security games. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '12, 1297–1298. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Robinson, D. J. 2010. *Cyber-based behavioral modeling*. Ph.D. Dissertation, Dartmouth College.

Tambe, M. 2011. *Security and game theory: Algorithms, deployed systems, lessons learned*. Cambridge University Press.

Zhou, Y., and Jiang, X. 2012. Dissecting android malware: Characterization and evolution. In *Security and Privacy (SP), 2012 IEEE Symposium on*, 95–109. IEEE.

Zhu, X. 2015. Machine teaching: an inverse approach to machine learning and an approach toward optimal education. AAAI.