Machine Learning in Adversarial Environments*

Pavel Laskov
University of Tübingen, Wilhelm-Schickard-Institute for Computer Science, Sand 1, 72070 Tübingen, Germany

Richard Lippmann
MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02173, USA

June 28, 2010

Abstract. Whenever machine learning is used to prevent illegal or unsanctioned activity and there is an economic incentive, adversaries will attempt to circumvent the protection provided. Constraints on how adversaries can manipulate training and test data for classifiers used to detect suspicious behavior make problems in this area tractable and interesting. This special issue highlights papers that span many disciplines including email spam detection, computer intrusion detection, and detection of web pages deliberately designed to manipulate the priorities of pages returned by modern search engines. The four papers in this special issue provide a standard taxonomy of the types of attacks that can be expected in an adversarial framework, demonstrate how to design classifiers that are robust to deleted or or corrupted features, demonstrate the ability of modern polymorphic engines to rewrite malware so it evades detection by current intrusion detection and antivirus systems, and provide approaches to detect web pages designed to manipulate web page scores returned by search engines. We hope that these papers and this special issue encouragesthe multidisciplinary cooperation required to address many interesting problems in this relatively new area including predicting the future of the arms races created by adversarial learning, developing effective long-term defensive strategies, and creating algorithms that can process the massive amounts of training and test data available for internet-scale problems.

Keywords: adversarial learning, adversary, spam, intrusion detection, web spam, robust classifier, feature deletion, arms race, game theory

1. Introduction

Machine learning techniques are increasingly used in environments where adversaries consciously act to limit or prevent accurate performance. A classical example is spam filtering where spammers tailor messages to avoid the most recent spam detection techniques. Further examples of adversarial learning arise in the field of computer security where an escalating arms race is taking place between detection and evasion techniques for various types of malware. In general, one can expect that whenever machine learning is used to provide protection against some illegal activity, adversaries will attempt to circumvent these approaches.

* This work is sponsored by the United States Air Force under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Vulnerability of machine learning methods to adversarial manipulation cannot be simply brushed off by a plea for new, "robust" methods. The theoretical foundations of machine learning are largely built on the assumption that training data adequately describes the underlying phenomena addressed by learning. This assumption is obviously violated if either the training or the test distributions are intentionally altered. Furthermore, even the consideration of *some* potential difference in the two distributions may not be appropriate. In the adversarial case, the learning is faced by a *worst-case* difference between the training and test distributions, as the attacker – unless bound by some problem-specific constraints – can be assumed to use any possible means to disrupt the learning algorithm. Learning methods for adversarial environments should therefore be designed to protect against malicious data distortion.

Protection against adversarial data may seem to be a "mission impossible". Indeed, an unconstrained adversary who can arbitrarily alter data and labels can induce an error rate of up to 100% (e.g. (Kearns and Li, 1993; Auer and Cesa-Bianchi, 1998; Bschouty et al., 2002)). In practice, however, an attacker must follow certain constraints. For example, a spam email must still deliver its message, malware sent to a host must correctly execute and exploit a vulnerability, and adversaries attempting to manipulate search engines can only control a fraction of all web sites. In some cases, it can be shown that constraints make finding an optimal attack computationally intractable (Fogla and Lee, 2006).

Investigation of learning methods for adversarial environments has been carried out in three largely distinct research areas: machine learning, computer security and spam filtering. In machine learning, previous work has centered around minimax methods with a goal of attaining robustness against input uncertainty. Robust classifiers have been developed to handle feature deletion (Globerson and Roweis, 2006) and general convex invariances (Teo et al., 2008). Researches have also demonstrated that unique Nash equilibia exist for some types of adversarial situations (Brückner and Scheffer, 2009). In addition, a theoretical analysis of robustness of certain learning algorithms against specific attacks has been carried out in recent work (Nelson et al., 2010; Kloft and Laskov, 2010).

A somewhat different view of adversarial learning problems has emerged in the field of computer security, especially intrusion detection. Several methods have been proposed for detecting anomalous network packets or to automatically generate signatures that are closely related to machine learning methods, e.g. one-class classification or naive Bayes (Wang and Stolfo, 2004; Wang et al., 2006; Newsome et al., 2005; Li et al., 2006). For all of these methods, effective attacks have been proposed shortly thereafter (Fogla et al., 2006; Perdisci et al., 2006) often followed again by more complex attack-resistant classifiers (Cretu et al., 2008). Although it has been shown that

intrusion detection systems that are difficult to spoof and that use "vulnerability signatures" can be created for malware that exploits known software vulnerabilities, these systems are often impractical and far too complex to protect large networks (Brumley et al., 2006).

Spam filtering is an area where we all have some experience with the problems of learning in adversarial environments. Various spam filter evasion techniques have not only been considered in the literature (Graham-Cumming, 2004; Wittel and Wu, 2004; Lowd and Meek, 2005b) but also witnessed by millions of email users. The urgency of spam filter evasion has lead to a significant interest to investigation of spam evasion constraints and robust spam filtering techniques (Dalvi et al., 2004; Lowd and Meek, 2005a).

The heterogeneity of previous work on learning methods for adversarial environments clearly calls for summarization of knowledge from various fields to establish a multidisciplinary dialogue between research communities. A starting point for this endeavor was the workshop "Machine Learning in Adversarial Environments for Computer Security" organized by the editors of this issue in December 2007 at the Neural Information Processing Systems – Natural and Synthetic (NIPS) conference (Lippmann and Laskov, 2007). The goal of this special issue is to deepen the common understanding of security issues arising in machine learning attained at that workshop via the presentation of recent novel work in this field.

2. Contributions of the Special Issue

The articles selected for this special issue reflect the diversity of scientific methodology in the various application domains of adversarial learning. Each submitted article was carefully reviewed by one member of the editorial board of the Machine Learning Journal and at least one external reviewer from a respective field. Although we did not have any quotas in mind, the selected articles span all important application domains. We hope that the presented work provides insights to a wide spectrum of interested readers. The following briefly introduces contributed articles.

The special issue begins with a discussion of fundamental issues related to the security of machine learning carried out by Barreno et al. (Barreno et al., 2010). The authors propose a taxonomy of attacks against machine learning algorithms inspired by the classical security goals, such as availability and integrity. A number of well-known attacks against machine learning algorithms are shown to belong to the specific categories in this taxonomy. As a second contribution, the authors propose a game-theoretic framework for the analysis of learning algorithms in adversarial environments which is suitable for formal specification of various attacks in their taxonomy. As an illustration of

their taxonomy, the authors show how it can guide the development of attacks against SpamBayes, a popular spam filter.

Dekel et al. (Dekel et al., 2010) apply machine learning techniques to create two-class linear classifiers designed to perform well when an adversary can corrupt or delete a constrained number of input features. They create a linear-programming solution and a more practical online perceptron-like algorithm and provide generalization bounds for these two approaches. Both attempt to create a robust classifier that uses many input features instead of focusing on a few important features that may be the first an adversary deletes. These new algorithms always performed as well as or better than a conventional linear SVM classifier. They outperformed the SVM classifier and other robust classifiers in handwritten digit recognition applications where there are many input features and a high level of feature redundancy.

The challenges to be faced by machine learning algorithms in the domain of malware analysis are investigated in (Song et al., 2010). The main motivation of this work is the exploding variability of malicious programs observed by security experts in the "wild". Unlike typical machine learning literature, the authors try to analyze the difficulty of the malware detection problem rather than propose a specific solution for it. The authors investigate various automated evasion techniques that enable malware writers to generate highly variable polymorphic versions of malware that all exploit the same software vulnerability. Two quantitative measures are proposed for evaluation of the strength of polymorphic engines: the variation strength and the propagation strength. Using these measures, the authors analyze variability of real shellcode examples and claim that the degree of variability attainable by polymorphic engines raises a strong doubt that attacks can ever me modeled by the simple generative approach (i.e. attack signatures) used in many common intrusion detection and antivirus tools.

The last article of this issue (Abernethy et al., 2010) goes back to a specific application scenario of learning in adversarial environments: web spam detection. Web spam poses a serious threat to web information retrieval. Its main goal is to skew the results of search or ranking queries by malicious content manipulation. A new algorithm WITCH for web spam detection is presented which simultaneously explores the structure of link information in web pages as well as content features. The method is efficient, scalable and provides a significant accuracy improvement on a standard web spam benchmark data.

3. Conclusions

Machine learning can provide solutions to difficult security problems on the internet included filtering spam email, detecting various types of attacks against servers and personal computers, and detecting web pages deliberately de-

signed to manipulate the web page priorities computed by modern search engines. All of these problems involve manipulation of massive amounts of data in a highly variable environment and the need for rapid and accurate training and classification. The existence of adversaries who can make a profit in these areas complicates the application of machine learning and creates an arms-race between those developing improved classifiers and adversaries trying to manipulate these classifiers. Fortunately, each application area provides constraints on adversarial actions that make classification, to some extent, feasible. Papers in this special issue provide an initial taxonomy of adversarial attacks and a sampling of approaches used by adversaries to defeat current methods and by defenders to create more robust classifiers. It is hoped that these papers will encourage multidisciplinary work that can focus on difficult problems including predicting the future of these types of arms races, developing effective long-term defensive strategies, developing algorithms that scale up to the massive amounts of data required to train and test systems across the internet, and incorporating various types of diversity and randomness in classifier training and use to prevent adversaries from predicting classifier outcomes.

References

Abernethy, J., O. Chapelle, and C. Castillo: 2010, 'Graph regularization methods for Web spam detection'. *Machine Learning Journal* **81**(2). DOI: 10.1007/s10994-009-5154-2.

Auer, P. and N. Cesa-Bianchi: 1998, 'On-line learning with malicious noise and the closure algorithm'. *Annals of Mathematics and Artificial Intelligence* **23**(1-2), 83–99.

Barreno, M., B. Nelson, A. Joseph, and D. Tygar: 2010, 'The security of machine learning'. *Machine Learning Journal* **81**(2). DOI: 10.1007/s10994-010-5188-5.

Brückner, M. and T. Scheffer: 2009, 'Nash Equilibria of Static Prediction Games'. In: Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta (eds.): *Advances in Neural Information Processing Systems 22*. pp. 171–179.

Brumley, D., J. Newsome, D. Song, H. Wang, and S. Jha: 2006, 'Towards automatic generation of vulnerability-based signatures'. In: *IEEE Symposium on Security and Privacy*. pp. 2–16.

Bschouty, N. H., N. Eiron, and E. Kushilevitz: 2002, 'PAC Learning with Nasty Noise'. *Theoretical Computer Science* **288**(3), 255–275.

Cretu, G., A. Stavrou, M. Locasto, S. Stolfo, and A. Keromytis: 2008, 'Casting out Demons: Sanitizing Training Data for Anomaly Sensors'. In: *IEEE Symposium on Security and Privacy*. pp. 81–95.

Dalvi, N., P. Domingos, M. Sumit, and S. D. Verma: 2004, 'Adversarial Classification'. In: *Knowledge Discovery in Databases*. pp. 99–108, ACM Press.

Dekel, O., O. Shamir, and L. Xiao: 2010, 'Learning to classify with missing and corrupted features'. *Machine Learning Journal* **81**(2). DOI: 10.1007/s10994-009-5124-8.

Fogla, P. and W. Lee: 2006, 'Evading network anomaly detection systems: formal reasoning and practical techniques'. In: *ACM Conference on Computer and Communications Security (CCS)*. pp. 59–68.

Fogla, P., M. Sharif, R. Perdisci, O. Kolesnikov, and W. Lee: 2006, 'Polymorphic Blending Attacks'. In: *USENIX Security Symposium*. pp. 241–256.

Globerson, A. and S. Roweis: 2006, 'Nightmare at Test Time: Robust Learning by Feature Deletion'. In: *International Conference on Machine Learning (ICML)*. pp. 353–360.

Graham-Cumming, J.: 2004, 'How to beat an Adaptive Spam Filter'. In: *MIT Spam Conference*.

Kearns, M. and M. Li: 1993, 'Learning in the presence of malicious errors'. *SIAM Journal on Computing* **22**(4), 807–837.

Kloft, M. and P. Laskov: 2010, 'Online Anomaly Detection under Adversarial Impac'. In: *13th International Conference on Artificial Intelligence and Statistics (AISTATS)*. pp. 405–412.

Li, Z., M. Sandhi, Y. Chen, M.-Y. Kao, and B. Chavez: 2006, 'Hamsa: fast signature generation for zero-day polymorphic worms with provable attack resilience'. In: *IEEE Symposium on Security and Privacy*. pp. 32–47.

Lippmann, R. and P. Laskov: 2007, 'NIPS 2007 Workshop on Machine Learning in Adversarial Environments for Computer Security, http://mls-nips07.first.fraunhofer.de/'.

Lowd, D. and C. Meek: 2005a, 'Adversarial Learning'. In: *Conference on Email and Anti-Spam*.

Lowd, D. and C. Meek: 2005b, 'Good Word Attacks on Statistical Spam Filters'. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 641–647.

Nelson, B., B. Rubinstein, L. Huang, A. Joseph, S. Lau, S. Lee, S. Rao, A. Tran, and D. Tygar: 2010, 'Near-Optimal Evasion of Convex-Inducing Classifiers'. In: *13th International Conference on Artificial Intelligence and Statistics (AISTATS)*. pp. 549–556.

Newsome, J., B. Karp, and D. Song: 2005, 'Polygraph: Automatically generating signatures for polymorphic worms'. In: *IEEE Symposium on Security and Privacy*. pp. 120–132.

Perdisci, R., D. Dagon, W. Lee, P. Fogla, and M. Sharif: 2006, 'Misleading Worm Signature Generators Using Deliberate Noise Injection'. In: *IEEE Symposium on Security and Privacy*. pp. 17–31.

Song, Y., M. E. Locasto, A. Stavrou, A. D. Keromytis, and S. J. Stolfo: 2010, 'On the infeasibility of modeling polymorphic shellcode'. *Machine Learning Journal* **81**(2). DOI: 10.1007/s10994-009-5143-5.

Teo, C. H., A. Globerson, S. Roweis, and A. Smola: 2008, 'Convex Learning with Invariances'. In: J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.): *Advances in Neural Information Processing Systems 20*. Cambridge, MA, pp. 1489–1496, MIT Press.

Wang, K., J. J. Parekh, and S. J. Stolfo: 2006, 'ANAGRAM: A Content Anomaly Detector Resistant To Mimicry Attack'. In: *Recent Adances in Intrusion Detection (RAID)*. pp. 226–248.

Wang, K. and S. Stolfo: 2004, 'Anomalous Payload-Based Network Intrusion Detection'. In: *Recent Adances in Intrusion Detection (RAID)*. pp. 203–222.

Wittel, G. and S. Wu: 2004, 'On Attacking Statistical Spam Filters'. In: *Conference on Email and Anti-Spam*.