Statistical Classification and Computer Security

Alvaro A. Cárdenas **J.D. Tygar**
University of California, Berkeley
{cardenas,tygar}@cs.berkeley.edu

1 Introduction

During the last couple of years we have been addressing classification problems in computer security applications. In particular, we have focused on three key items: (1) evaluation metrics for classifiers in adversarial environments, (2) the design of optimal classifiers against adaptive adversaries, and (3) voting algorithms for the combination of multiple classifiers. Our results have been published in computer security [1, 2, 3, 4] and machine learning [5, 6, 7, 8] venues. In this abstract we summarize our results, discuss our current work, and mention some other open problems in the intersection of machine learning and computer security.

2 Evaluation Metrics

Measuring the classification performance of Intrusion Detection Systems (IDS) with typical metrics such as the *accuracy* of the classifier or the *Receiver Operating Characteristic (ROC) curves* tend to give results that are easily misinterpreted. The root cause of this problem is the large *class imbalance* between the normal events and attack events [9]. To alleviate this problem, the IDS community proposed several other metrics, such as the *Bayesian detection rate* [9], the *expected cost* [10], the *sensitivity* [11], and the *intrustion detection capability* [12].

Despite making their own contributions to the problem, these metrics were introduced by different theoretical backgrounds, were difficult to compare to each other, and assumed the knowledge of uncertain parameters.

In an effort to alleviate these problems we (1) introduced a unified view of the metrics by casting the intrusion evaluation problem as a multi-criteria optimization problem, (2) proposed a new evaluation metric called the IDOC curves[1] and, (3) considered the least favorable uncertain parameters to obtain a robust evaluation of the system [1]. Some of our results were relevant not only to the IDS community but also to machine learning problems with large class imbalances [6].

We believe there are several extensions to the evaluation of classifiers used in computer security, and in particular, to understanding the tradeoffs involved between utility and security [13]. For example, formal problems in security are often *undecidable* (such as the detection of malicious code), and therefore most formal algorithms usually settle down for either *soundness* or *completeness* (but not both). The tradeoffs for loosing "some" soundness for the benefit of obtaining "more" completeness or vice-versa, have not been studied. Another example comes from finance, and involves the economic analysis of investments for security technologies. To fully understand the risk model, metrics such as the *expected loss* might not be enough in some cases, and we require more information about the distribution of the losses (such as the variance of the losses).

3 Designing Optimal Classifiers Against Adaptive Adversaries

The main problem in computer security is the presence of an active adversary that tries to evade or attack any security mechanism in place. While the security literature has developed a culture of analyzing and formalizing several adversary models (such as the Dolev-Yao threat model [14]), the modeling of threats in the machine learning community is still in its infancy; in particular, several

[1] IDOC curves were later renamed B-ROC curves [6].

approaches still assume that the attacker will behave similarly before and after the deployment of the classifier!

To address this problem we have been developing a series of *adaptive* threat models against statistical classification systems in computer security, and using these threat models to design robust classification systems [5]. Our formulations are essentially a *game-theoretic* interpretation of the classification problem. Let Ψ denote the metric that our classifier wants to minimize. Let $\mathcal{C}$ denote a set of possible classifiers, and let $\mathcal{A}$ denote our adversary model. Because an adaptive adversary always makes the final move in the game, we need to solve the following equation:

$$\min_{C \in \mathcal{C}} \max_{A \in \mathcal{A}} \Psi(C, A). \tag{1}$$

A way to show that (C^*, A^*) satisfies Eq. (1) is to show that the pair (C^*, A^*) forms a saddle point equilibrium:

$$\forall (C, A) \in \mathcal{C} \times \mathcal{A} \quad \Psi(C^*, A) \leq \Psi(C^*, A^*) \leq \Psi(C, A^*). \tag{2}$$

Our first formulation following these principles considered a simple black-box adversary model for IDS that achieved better properties that IDS whose parameters were set without considering future threats [1]. We then developed more complex adversary models for detecting misbehavior in the Medium Access Control (MAC) layer of wireless networks [2]. One of our main results [3] shows how our classification algorithms outperform previously proposed MAC-layer detection algorithms that do not consider threat models. Finally, we have also used our game-theoretic formulation of classification for multimedia watermarking [15, 8].

It is important to point out that we've had success designing classifiers resilient to adaptive adversaries because we have carefully selected problems and abstractions that are simple enough to remain tractable, yet meaningful enough for useful solutions. There are, however, several other problems that might not be amenable to simple abstractions, and the main limitation will be the tractability of the game-theoretic solution. We believe there are still several open problems in trying to come up with useful threat models for classifier systems, considering the tradeoffs between multiple metrics, and in investigating the theoretical limitations of several, more complex, formulations.

4 Combination of Classifiers

By using traditional hypothesis testing theory we were able to explain and interpret in a new theoretical light, previous work in the machine learning community for combining classifiers, and at the same time, propose a new *stacking* algorithm that outperforms previous approaches [7].

We have been working on practical applications of this algorithm for combining the alarms of different IDS [4]. We believe this is an example of how principled and theoretically sound approaches to security can be derived from machine learning techniques.

Combining classifiers is an active problem in several computer security scenarios where diversity and redundancy are used to increase the reliability of systems, or against *Byzantine* adversaries [16] (or malicious insiders), where an entity has to make a security decision based on conflicting pieces of information. We believe that these problems can learn many useful techniques from the machine learning community.

5 Open Problems and Future Work

There are many other fields where the combination of machine learning and computer security can be mutually beneficial. Two of these fields are *cryptanalysis* and *traffic analysis*.

Machine learning and cryptanalysis are sister fields because the goal of the cryptanalyst is to learn the unknown function (the decryption function) from examples of its input/output behavior, and prior knowledge about the class of possible functions [17, 18]. Cryptography, on the other hand, can be used to show that certain learning problems are computationally intractable. This problems can help us understand the theoretical limits of machine learning.

Although theoretically interesting, cryptanalysis might not have a big impact in the practice of computer security because most well-designed cryptographic algorithms are hard to break. However, we believe traffic analysis is a major open field where machine learning can be used to extract side information from encrypted communications. Examples of information leaked can be the source or destination of the communication, the type of protocols used, social interactions, stepping stones, and many more [19]. We are currently working in this field and we expect to have results to show in the workshop.

References

[1] Alvaro A. Cárdenas, John Baras, and Karl Seamon. A framework for the evaluation of intrusion detection systems. In *IEEE Symposium on Security & Privacy (S&P '06)*, pages 63–77, Berkeley/Oakland, California, May 2006.

[2] Svetlana Radosavac, Alvaro A. Cárdenas, John S. Baras, and George Moustakides. Detecting IEEE 802.11 MAC layer misbehavior in ad hoc networks: Robust strategies against individual and colluding attackers. In *Journal of Computer Security*, volume 15, pages 103–128. IOS Press, January 2007.

[3] Alvaro A. Cárdenas, Svetlana Radosavac, and John S. Baras. Performance comparison of detection schemes for MAC layer misbehavior. In *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM 2007)*, pages 1496–1504, May 2007.

[4] Guofei Gu, Alvaro A. Cárdenas, and Wenke Lee. Principled reasoning and practical applications of alert fusion in intrusion detection systems. In *Submitted to the ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, 2008.

[5] Alvaro A. Cárdenas and John S. Baras. Evaluation of classifiers and learning rules: Considerations for security applications. In *Evaluation Methods for Machine Learning, Papers from the AAAI Workshop*, pages 30–35. AAAI Press, July 17 2006.

[6] Alvaro A. Cárdenas and John S. Baras. B-ROC curves for the assesment of classifiers over imbalanced data sets. In *Proceedings of the twenty-first National Conference on Artificial Intelligence, (AAAI 06)*, Boston, Massachusetts, July 16–20 2006.

[7] Marco Barreno, Alvaro A. Cárdenas, and J. D. Tygar. Optimal ROC curve for a combination of classifiers. In *Advances in Neural Information Processing Systems (NIPS 2007)*, December 2007.

[8] Alvaro A. Cárdenas, George Moustakides, and John S. Baras. On optimal watermarking schemes in uncertain gaussian channels. In *Proceedings of the 2007 IEEE International Conference on Image Processing (ICIP' 07)*, September 15–20 2007.

[9] Stefan Axelsson. The base-rate fallacy and its implications for the difficulty of intrusion detection. In *Proceedings of the 6th ACM Conference on Computer and Communications Security (CCS '99)*, pages 1–7, November 1999.

[10] John E. Gaffney and Jacob W. Ulvila. Evaluation of intrusion detectors: A decision theory approach. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, pages 50–61, Oakland, CA, USA, 2001.

[11] Giovanni Di Crescenzo, Abhrajit Ghosh, and Rajesh Talpade. Towards a theory of intrusion detection. In *ESORICS 2005, 10th European Symposium on Research in Computer Security*, pages 267–286, Milan, Italy, September 12-14 2005. Lecture Notes in Computer Science 3679 Springer.

[12] Guofei Gu, Prahlad Fogla, David Dagon, Wenke Lee, and Boris Skoric. Measuring intrusion detection capability: An information-theoretic approach. In *Proceedings of ACM Symposium on InformAtion, Computer and Communications Security (ASIACCS '06)*, Taipei, Taiwan, March 2006.

[13] Alvaro A. Cárdenas, Gelareh Taban, and John S. Baras. A unified framework of information assurance for the design and analysis of security algorithms. In *Proceedings of the 25th Army Science Conference*, Orlando, FL, November 27–30 2006.

[14] D. Dolev and A. Yao. On the security of public key protocols. *IEEE Transactions on Information Theory*, 29(2):198–208, March 1983.

[15] Alvaro A. Cárdenas, George Moustakides, and John S. Baras. Towards optimal design of data hiding algorithms against nonparametric adversary models. In *41st Annual Conference on Information Sciences and Systems (CISS'07)*, pages 911–916. IEEE Press, March 14–16 2007.

[16] Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, July 1982.

[17] Ron Rivest. Cryptography and machine learning. In *Advances in Cryptology – ASIACRYPT '91*, volume 739 of *Lecture Notes in Computer Science*, pages 427–439. Springer Berlin / Heidelberg, 1993.

[18] Joshua Mason, Kathryn Watkins, Jason Eisner, and Adam Stubblefield. A natural language approach to automated cryptanalysis of two-time pads. In *Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS 06)*, pages 235–244, October 2006.

[19] George Danezis and Richard Clayton. Introducing traffic analysis. *Digital Privacy: Theory, Technologies and Practices*, December 2007.