Machine Learning under Attack: Vulnerability Exploitation and Security Measures

Invited Keynote

Battista Biggio
Department of Electrical and Electronic Engineering
University of Cagliari, Piazza d'Armi, 09123 Cagliari, Italy
battista.biggio@diee.unica.it

ABSTRACT

Learning to discriminate between secure and hostile patterns is a crucial problem for species to survive in nature. Mimetism and camouflage are well-known examples of evolving weapons and defenses in the arms race between predators and preys. It is thus clear that all of the information acquired by our senses should not be considered necessarily secure or reliable. In machine learning and pattern recognition systems, however, we have started investigating these issues only recently. This phenomenon has been especially observed in the context of adversarial settings like malware detection and spam filtering, in which data can be purposely manipulated by humans to undermine the outcome of an automatic analysis. As current pattern recognition methods are not natively designed to deal with the intrinsic, adversarial nature of these problems, they exhibit specific vulnerabilities that an attacker may exploit either to mislead learning or to evade detection. Identifying these vulnerabilities and analyzing the impact of the corresponding attacks on learning algorithms has thus been one of the main open issues in the novel research field of adversarial machine learning, along with the design of more secure learning algorithms.

In the first part of this talk, I introduce a general framework that encompasses and unifies previous work in the field, allowing one to systematically evaluate classifier security against different, potential attacks. As an example of application of this framework, in the second part of the talk, I discuss evasion attacks, where malicious samples are manipulated at test time to evade detection. I then show how carefully-designed poisoning attacks can mislead some learning algorithms by manipulating only a small fraction of their training data. In addition, I discuss some defense mechanisms against both attacks in the context of real-world applications, including biometric identity recognition and computer security. Finally, I briefly discuss our ongoing work on attacks against clustering algorithms, and sketch some promising future research directions.

IH&MMSec 2016 June 20-23, 2016, Vigo, Spain

ACM ISBN 978-1-4503-4290-2/16/06.

DOI: http://dx.doi.org/10.1145/2909827.2930784

CCS Concepts

•**Security and privacy** → *Malware and its mitigation; Denial-of-service attacks;* Biometrics; •**Computing methodologies** → **Machine learning; Adversarial learning;**

Keywords

Secure Pattern Recognition; Adversarial Machine Learning; Evasion Attacks; Poisoning Attacks

Short Biography

Battista Biggio received the M.Sc. degree (with honors) in Electronic Engineering (2006) and the Ph.D. degree in Electronic Engineering and Computer Science (2010) from the University of Cagliari (Italy). Since 2007, he has been with the Department of Electrical and Electronic Engineering of the University of Cagliari, where he is currently a post-doctoral researcher. In 2011, he visited the University of Tübingen (Germany), and worked on the security of learning algorithms to training data contamination. His research interests include secure machine learning, multiple classifier systems, kernel methods, computer security and biometrics. On these topics, he has published more than 50 papers on international conferences and journals, collaborating with several research groups from academia and companies throughout the world. Dr. Biggio has also recently co-founded a company named Pluribus One, where he is responsible of leveraging machine-learning algorithms to drive product innovation. He regularly serves as a reviewer and program committee member for several international conferences and journals on the aforementioned research topics. Dr. Biggio is a member of the IEEE and of the IAPR.

1. REFERENCES

[1] M. Barreno, B. Nelson, A. Joseph, and J. Tygar. The security of machine learning. *Machine Learning*, 81:121–148, 2010.

[2] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar. Can machine learning be secure? In *ACM Symp. Information, Computer and Comm. Sec.*, ASIACCS '06, pages 16–25, New York, NY, USA, 2006. ACM.

[3] B. Biggio, S. R. Bulò, I. Pillai, M. Mura, E. Z. Mequanint, M. Pelillo, and F. Roli. Poisoning complete-linkage hierarchical clustering. In P. Franti, G. Brown, M. Loog, F. Escolano, and M. Pelillo,

editors, *Joint IAPR Int'l Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, volume 8621 of *LNCS*, pages 42–52, Joensuu, Finland, 2014. Springer Berlin Heidelberg.

[4] B. Biggio, I. Corona, Z.-M. He, P. P. K. Chan, G. Giacinto, D. S. Yeung, and F. Roli. One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time. In F. Schwenker, F. Roli, and J. Kittler, editors, *Multiple Classifier Systems*, volume 9132 of *LNCS*, pages 168–180. Springer International Publishing, 2015.

[5] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, editors, *European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Part III*, volume 8190 of *LNCS*, pages 387–402. Springer Berlin Heidelberg, 2013.

[6] B. Biggio, I. Corona, B. Nelson, B. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli. Security evaluation of support vector machines in adversarial environments. In Y. Ma and G. Guo, editors, *Support Vector Machines Applications*, pages 105–153. Springer International Publishing, 2014.

[7] B. Biggio, L. Didaci, G. Fumera, and F. Roli. Poisoning attacks to compromise face templates. In *6th IAPR Int'l Conf. on Biometrics (ICB 2013)*, pages 1–7, Madrid, Spain, 2013.

[8] B. Biggio, G. Fumera, and F. Roli. Multiple classifier systems for robust classifier design in adversarial environments. *Int'l J. Mach. Learn. and Cybernetics*, 1(1):27–41, 2010.

[9] B. Biggio, G. Fumera, and F. Roli. Pattern recognition systems under attack: Design issues and research challenges. *Int'l J. Patt. Recogn. Artif. Intell.*, 28(7):1460002, 2014.

[10] B. Biggio, G. Fumera, and F. Roli. Security evaluation of pattern classifiers under attack. *IEEE Trans. on Knowledge and Data Engineering*, 26(4):984–996, April 2014.

[11] B. Biggio, G. Fumera, F. Roli, and L. Didaci. Poisoning adaptive biometric systems. In G. Gimel'farb, E. Hancock, A. Imiya, A. Kuijper, M. Kudo, S. Omachi, T. Windeatt, and K. Yamada, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 7626 of *LNCS*, pages 417–425. Springer Berlin Heidelberg, 2012.

[12] B. Biggio, G. Fumera, P. Russu, L. Didaci, and F. Roli. Adversarial biometric recognition : A review on biometric system security from the adversarial machine-learning perspective. *Signal Processing Magazine, IEEE*, 32(5):31–41, Sept 2015.

[13] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In J. Langford and J. Pineau, editors, *29th Int'l Conf. on Machine Learning*, pages 1807–1814. Omnipress, 2012.

[14] B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, and F. Roli. Is data clustering in adversarial settings secure? In *Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security*, AISec '13, pages 87–98, New York, NY, USA, 2013. ACM.

[15] M. Brückner, C. Kanzow, and T. Scheffer. Static prediction games for adversarial learning problems. *J. Mach. Learn. Res.*, 13:2617–2654, September 2012.

[16] D. M. Freeman, S. Jain, M. Dürmuth, B. Biggio, and G. Giacinto. Who are you? a statistical approach to measuring user authenticity. In *23rd Annual Network & Distributed System Security Symposium (NDSS)*. The Internet Society, 2016.

[17] L. Huang, A. D. Joseph, B. Nelson, B. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *4th ACM Workshop on Artificial Intelligence and Security (AISec 2011)*, pages 43–57, Chicago, IL, USA, 2011.

[18] M. Kloft and P. Laskov. Online anomaly detection under adversarial impact. In *Proceedings of the 13th Int'l Conf. on Artificial Intelligence and Statistics (AISTATS)*, pages 405–412, 2010.

[19] M. Kloft and P. Laskov. Security analysis of online centroid anomaly detection. *Journal of Machine Learning Research*, 13:3647–3690, 2012.

[20] D. Maiorca, I. Corona, and G. Giacinto. Looking at the bag is not enough to find the bomb: an evasion of structural methods for malicious pdf files detection. In *8th ACM SIGSAC Symp. on Information, Computer and Comm. Sec.*, ASIA CCS '13, pages 119–130, New York, NY, USA, 2013. ACM.

[21] F. Roli, B. Biggio, and G. Fumera. Pattern recognition systems under attack. In J. Ruiz-Shulcloper and G. S. di Baja, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 8258 of *LNCS*, pages 1–8. Springer, 2013.

[22] B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, and J. D. Tygar. Antidote: understanding and defending against poisoning of anomaly detectors. In *9th ACM SIGCOMM Internet Measurement Conf.*, IMC '09, pages 1–14, New York, NY, USA, 2009. ACM.

[23] N. Šrndić and P. Laskov. Detection of malicious pdf files based on hierarchical document structure. In *20th Annual Network & Distributed System Security Symposium (NDSS)*. The Internet Society, 2013.

[24] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli. Is feature selection secure against training data poisoning? In F. Bach and D. Blei, editors, *JMLR W&CP - Proc. 32nd Int'l Conf. Mach. Learning (ICML)*, volume 37, pages 1689–1698, 2015.

[25] F. Zhang, P. Chan, B. Biggio, D. Yeung, and F. Roli. Adversarial feature selection against evasion attacks. *IEEE Trans. on Cybernetics*, 46(3):766–777, 2016.