Poster: A Modular and Innovative Security Analytics Course

Rakesh M. Verma

Comp. Sci. Dept., University of Houston, Texas
rverma@uh.edu

ABSTRACT

Techniques from data science are increasingly being applied by researchers to security challenges. However, requirements unique to the security domain necessitate painstaking care for the models to be valid and robust. In this paper, we outline a novel security analytics course, its modular design, some of its key innovations, and experience in teaching it.

1 INTRODUCTION/PROBLEM

Data Science techniques are increasingly being applied to security challenges such as malware detection, intrusion detection, and phishing [1, 2], among others. However, a serious problem in the cybersecurity domain is the presence of an active attacker, who is trying to defeat the methods being deployed. There are other issues also, such as: availability and quality of datasets, the base-rate fallacy, time-scale of attacks, asymmetrical costs of misclassification, and potential data poisoning [7, 9].

For these reasons, data science techniques need to be *adapted*, as opposed to directly applied, to security challenges. Successful adaptation of techniques from one field into another requires professionals, who share a common vocabulary and understanding. This is the goal of our new course, which has been approved as a regular course at both senior undergraduate and graduate levels.

1.1 Background/Related Work

A 10-course graduate program with security and analytics is in [5], a case study of introducing data analytics in an advanced cybersecurity course is in [4], and some universities (e.g., Stanford/UT Dallas) have a data mining for cybersecurity course. Most closely related to our work is [3]. However, they do not consider the unique security needs and adaptation, and the design is not modular.

2 OVERVIEW/METHODS/RESULTS

At the Univ. of Houston, an NSA/DHS center of academic excellence in information assurance/cyberdefense education and research, I have been teaching a security analytics course annually since 2015. The course design is modular, consisting of four modules: 1) basics of security, 2) data mining for security (data handling & mostly unsupervised methods), 3) machine learning methods (mainly supervised/semi-supervised) for security, and 4) text mining/natural language processing (NLP) for security [8].

Each module is about 3-4 weeks long and consists of three hours of lectures, which are supplemented by conceptual and laboratory

exercises, a pre-test, a post-test, and a quiz. Since the emphasis is on techniques for security, each module contains a discussion of security challenges that have been addressed using the methods. For example, clustering in Module 2 is illustrated with malware clustering, and NLP techniques such as n-grams of characters and Part-of-Speech tagging in Module 4 are illustrated with password security and phishing email detection. General principles, approaches, and in-depth study are emphasized, since we want the students to innovatively adapt the techniques rather than “black-box” apply the techniques. A project with real-world dataset is a significant component of the course. Keeping the domain of security in focus, we introduce techniques for imbalanced datasets, semi-supervised methods, metrics for evaluating models on imbalanced datasets, adversarial machine learning, cost-based classification and evaluation, online learning, and learning methods for nonstationary data [6]. **Preliminary Results.** We are currently conducting quantitative analysis of the results, but qualitative surveys already indicate that the course is being appreciated by students. For example, a graduate student said that they learned more about data mining and machine learning techniques in this course than they learned in a data mining/machine learning course.

3 CONTRIBUTIONS AND FUTURE WORK

We have designed and implemented a new course integrating useful concepts from data science/analytics and cybersecurity. Our course is made up of four “plug-and-play” modules, which can be adapted for other courses. We continue to revise the course materials (available online) and future versions of the course will also include deeper discussions of data quality in security challenges.

Acknowledgments

Thanks to NSF for partial support under grants DUE 1241772, CNS 1319212, DGE 1433817, and DUE 1356705.

REFERENCES

- [1] A. Das, S. Baki, A. El Aassal, R. Verma, and A. Dunbar. 2019. SoK: A Comprehensive Reexamination of Phishing Research from the Security Perspective. *IEEE Communications Surveys Tutorials* (2019). <https://doi.org/10.1109/COMST.2019.2957750>
- [2] G. Egozi and R. Verma. 2018. Phishing Email Detection Using Robust NLP Techniques. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. 7–12. <https://doi.org/10.1109/ICDMW.2018.00009>
- [3] A. Epishkina and S. Zapechnikov. 2016. A syllabus on data mining and machine learning with applications to cybersecurity. In *3rd IEEE DIPDMWC*. 194–199.
- [4] Basil Hamdan. 2017. Teaching Case Study: Introducing Data Analytics in an Advanced Cybersecurity Course. *J. Comput. Sci. Coll.* 33, 2 (Dec. 2017), 113–120.
- [5] S.A. Kumar. 2014. Designing a Graduate Program in Information Security and Analytics. In *Proc. 15th Ann. Conf. on Info. Tech. Edu. (SIGITE '14)*. ACM, 141–146.
- [6] R. Verma. 2018. Security Analytics: Adapting Data Science for Security Challenges. In *Proc. 4th ACM Int'l Workshop on Security & Privacy Analytics (IWSPA '18)*. 40–41.
- [7] R.M. Verma, M. Kantarcioglu, D.J. Marchette, E.L. Leiss, and T. Solorio. 2015. Security Analytics: Essential Data Analytics Knowledge for Cybersecurity Professionals and Students. *IEEE Security & Privacy* 13, 6 (2015), 60–65.
- [8] R.M. Verma and D.J. Marchette. 2019. *Cybersecurity Analytics*. CRC Press, Boca Raton, FL.
- [9] R.M. Verma, V. Zeng, and H. Faridi. 2019. Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets. In *Proc. 2019 ACM SIGSAC Conf. on Computer and Comm. Security (CCS '19)*. ACM, 2605–2607.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGCSE '20, March 11–14, 2020, Portland, OR, USA

© 2020 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6793-6/20/03.

<https://doi.org/10.1145/3328778.3372649>