ARTICLE N°6

A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection

Anna L.Buczak , *Member, IEEE*, and Erhan Guven, *Member, IEEE*

1. Introduction

1.1 Context

In a world where cyber security take his essort,this article comes at named the right moment to bring novelty of a literature survey of machine learning (ML) and data mining (DM) methods for cyber security applications. The ML/DM methods are described, as well as several applications of each method to cyber intrusion detection problems. The complexity of different ML/DM algorithms is discussed, and the paper provides a set of comparison criteria for ML/DM methods and a set of recommendations on the best methods to use depending on the characteristics of the cyber problem to solve.

1.2 Background

Cyber security is the set of technologies and processes designed to protect computers, networks, programs, and data from attack, unauthorized access, change, or destruction. Cyber security systems are composed of network security systems and computer (host) security systems. Each of these has, at a minimum, a firewall, anti-virus software, and an intrusion detection system (IDS). IDSs help discover, determine, and identify unauthorized use,duplication, alteration, and destruction of information systems [1]. The security breaches include external intrusions (attacks from outside the organization) and internal intrusions (attacks from within the organization).

Machine learning, field of study of the artificial intelligence, relates to the design, the analysis, the development and the implementation of the methods making it possible a machine(au direction large) to evolve/move by a systematic process, and thus to fill the spots difficult or problematic to fill by more traditional algorithmic means .

Data mining is a essential composite of technonologies big data and techniques of analyse the bulky data. It acts of the source of the big dated analytics, of the predictive analyses and the exploitation of the data .

This document based on the number of citations or the relevance of an emerging method, it representing each method were identified, read, and summarized. Because data are so important in ML/DM approaches, some well-known cyber data sets used in ML/DM are described. The complexity of ML/DM algorithms is addressed, discussion of challenges for using ML/DM for cyber security is presented, and some recommendations on when to use a given method are provided.

1.3 Problem and research questions

-how use of different machine learning and data mining technics for cyber domain ?
-Which are The main types of cyber analytics in support of intrusion detection system(IDS) ?
-how they function ?
-which is most efficient ?

2. Related works

TABLE : ML and DM METHODS and DATA they use

PAPERS	DATA USED	HYBRID	ANOMALY	MESUSE
J. Cannady, “Artificial neural networks for misuse detection,” in *Proc.1998 Nat. Inf. Syst. Secur. Conf.*, Arlington, VA, USA, 1998, pp. 443–456.	Network packet-level data			√
R. P. Lippmann and R. K. Cunningham, “Improving intrusion detection performance using keyword selection and neural networks,” *Comput. Netw.*, vol. 34, pp. 597–603, 2000.	DARPA 1998		√	
A. Bivens, C. Palagiri, R. Smith, B. Szymanski, and M. Embrechts,“Network-based intrusion detection using neural networks,” *Intell. Eng. Syst. Artif. Neural Netw.*,	DARPA 1999		√	

vol.12, no. 1,pp. 579–584				
H. Brahmi, B. Imen, and B. Sadok, “OMC-IDS: At the cross-roads of OLAP mining and intrusion detection,” in *Advances in Knowledge Discovery and Data Mining*. New York, NY, USA: Springer, 2012, pp. 13–24.	DARPA 1998			√
H. Zhengbing, L. Zhitang, and W. Junqi, “A novel network intrusion detection system (NIDS) based on signatures search of data mining,” in *Proc. 1st Int. Conf. Forensic Appl. Techn. Telecommun. Inf. Multimedia Workshop (e-Forensics ‘08)*, 2008, pp. 10–16.	Attack signatures(10 to 0 MB)			√
R. Hendry and S. J. Yang, “Intrusion signature creation via clustering anomalies,” in *Proc. SPIE Defense Secur. Symp. Int. Soc. Opt. Photonics*, 2008, pp. 69730C–69730C	KDD 1999			√
. M. Blowers and J. Williams, “Machine learning applied to cyber operations,” in *Network Science and CyBER ANAL*	KDD 1999		√	
N. B. Amor, S. Benferhat, and Z. Elouedi, “Naïve Bayes vs. Decision trees in intrusion detection systems,” in *Proc ACM Symp. Appl. Comput.*,2004, pp. 420–424.	KDD 1999		√	

Y. Hu and B. Panda, “A data mining approach for database intrusion detection,” in *Proc. ACM Symp. Appl. Comput.*, 2004, pp. 711-716.	DATA BASE LOGS			√
Z. Li, A. Zhang, J. Lei, and L. Wang, “Real-time correlation of network security alerts,” in *Proc. IEEE Int. Conf. e-Business Eng.*, 2007, pp. 73–80.	DARPA 2000		√	
Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, and K. Dai, “An efficient intrusion detection system based on support vector machines and gradually feature removal method,” *Expert Syst. Appl.*, vol. 39, no. 1, pp. 424–430,2012.	KDD 1999			√
F. Amiri, M. Mahdi, R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani,“Mutual information-based feature selection for IDSs,” *J. Netw. Comput. Appl.*, vol. 34, no. 4, pp. 1184–1199, 2011.	KDD 1999			√
W. J. Hu, Y. H. Liao, and V. R. Vemuri, “Robust support vector machines for anomaly detection in computer security,” in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 282–289.	DARPA 1998		√	
C. Wagner, F. Jérôme, and E. Thomas, “Machine learning approach forIP-flow record anomaly detection,” in *Networking 2011*. New York, NY, USA: Springer, 2011, pp. 28–39.	NET FLOW		√	

3. Methodology employed

For the realization of this work,the author employs a particular methodology .

They begin with the enumeration of different types of cyber analytics in support for intrusion detection .there are three main types : misuse-based (sometimes also called signature based), anomaly-based, and hybrid. Misuse-based techniques are designed to detect known attacks by using signatures of those attacks. They are effective for detecting known type of attacks without generating an overwhelming number of false alarms. They require frequent manual updates of the database with rules and signatures. Misuse-based techniques cannot detect novel (zero-day) attacks.

Anomaly-based techniques model the normal network and system behavior, and identify anomalies as deviations from normal behavior. They are appealing because of their ability to detect zero-day attacks. Another advantage is that the profiles of normal activity are customized for every system, application, or network, thereby making it difficult for attackers to know which activities they can carry out undetected. Additionally, the data on which anomaly-based techniques alert (novel attacks) can
be used to define the signatures for misuse detectors. The main disadvantage of anomaly-based techniques is the potential for high false alarm rates (FARs) because previously unseen (yet legitimate) system behaviors may be categorized as anomalies.

Hybrid techniques combine misuse and anomaly detection. They are employed to raise detection rates of known intrusions and decrease the false positive (FP) rate for unknown attacks. An in-depth review of the literature did not discover many pure anomaly detection methods; most of the methods were really hybrid. Therefore, in the descriptions of ML and DM methods, the anomaly detection and hybrid methods are described together.

In continuation the authors, focus on major steps in Machine Learning and Data Mining,discuss cyber security data sets used in Machine Learning and Data Mining .After,they describe the individual methods and related papers for Machine Learning and Data Mining in cyber security ; discus the computational complexity of different methods, describes observations and recommendations.
Lastly, they present conclusions.

4. Results found and Perspectives

The paper describes the literature review of ML and DM methods used for cyber. Special emphasis was placed on finding example papers that describe the use of different ML and DM techniques in the cyber domain, both for misuse and anomaly detection. Unfortunately, the methods that are the most effective for cyber applications have not been established; and given the richness and complexity of the methods, it is impossible to make one recommendation for each method, based on the type of attack the system is supposed to detect. When determining the effectiveness of the methods, there is not one criterion but several criteria that need to be taken into account.

They include (as described in Section VI, Subsection C) accuracy, complex, time for classifying an unknown instance with a trained model, and understandability of the final solution (classification) of each ML or DM method. Depending on the particularIDS, some might be more important than others. Another crucial aspect of ML and DM for cyber intrusion detection is the importance of the data sets for training and testing the systems.ML and DM methods cannot work without representative data, and it is difficult and time consuming to obtain such data sets. To be able to perform anomaly detection and misuse detection, it is advantageous for an IDS to be able to reach network- and kernel-level data.

5. Most important publications for this work.

- I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Mateo, CA, USA: Morgan Kaufmann, 2011.
- K. Sequeira and M. Zaki, “ADMIT: Anomaly-based data mining for intrusions,” in *Proc 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2002, pp. 386–395
- J. Zhang, M. Zulkernine, and A. Haque, “Random-forests-based network intrusion detection systems,” *IEEE Trans. Syst. Man Cybern. C:Appl. Rev.*, vol. 38, no. 5, pp. 649–659, Sep. 2008
- Snort 2.0. *Sourcefire* [Online]. Available: http://www.sourcefire.com/technology/whitepapers.htm, accessed on Jun. 2014
- R. Lippmann, J. Haines, D. Fried, J. Korba, and K. Das, “The 1999 DARPA offline intrusion detection evaluation,” *Comput. Netw.*, vol. 34, pp. 579–595, 2000.

- J. Cannady, “Artificial neural networks for misuse detection,” in *Proc. 1998 Nat. Inf. Syst. Secur. Conf.*, Arlington, VA, USA, 1998, pp. 443–456.
- A. Bivens, C. Palagiri, R. Smith, B. Szymanski, and M. Embrechts, “Network-based intrusion detection using neural networks,” *Intell. Eng. Syst. Artif. Neural Netw.*, vol. 12, no. 1, pp. 579–584, 2002.
- C. Livadas, R. Walsh, D. Lapsley, and W. Strayer, “Using machine learning techniques to identify botnet traffic,” in *Proc 31st IEEE Conf. Local Comput. Netw.*, 2006, pp. 967–974.
- F. Jemili, M. Zaghdoud, and A. Ben, “A framework for an adaptive intrusion detection system using Bayesian network,” in *Proc. IEEE Intell. Secur. Informat.*, 2007, pp. 66–70.
- J. Hansen, P. Lowry, D. Meservy, and D. McDonald, “Genetic programming for prevention of cyberterrorism through dynamic and evolving intrusion detection,” *Decis. Support Syst.*, vol. 43, no. 4, pp. 1362–1374, Aug. 2007.

LINKS :

- https://en.wikipedia.org/wiki/Machine_learning 18/03/17.
- https://en.wikipedia.org/wiki/data_mining 16/03/17.

__How to write the background section of a simple research article

The background section is important.

- You need to introduce the topic;
- teach the reader about your topic and provide lots of interesting information;
- state the issue or controversy; and state why you are doing the research.