UNIVERSITE DE NGAOUNDERE

Faculté des Sciences

Département de Mathématiques et Informatique

UNIVERSITY OF NGAOUNDERE

Faculty of Science

Department of Mathematics and Computer Science

Report of the study of article 6, namely **« A survey of Data Mining and Machine Learning Methods for cyber security intrusion detection »,** whose the authors are :Anna L. Buczak, Member, IEEE, and Erhan Guven, Member, IEEE

By

kengne kengne joel 13A054FS
Mohamed Ali

conducted by

Dr.-ing Franklin Tchakounté

Academic year : 2016/2017

Table of content page

1.INTRODUCTION

1.1.Context

The security has become a great factor when designing of informations systems today, mostly in cyber applications because more and more, day after day different types of attacks take life and increasingly grow. The Reason of this growing is the rapid development of business, then a lot of data is online and has to be secured from any type of attackers. This paper focuses its study on the use of methods employed in machine learning and data mining for to optimize the cyber security intrusion detection with an emphasis on ML/DM methods and their descriptions .

So, before to go ahead let's define some terms which will help us to understand the problem in this survey .

An intrusion is an assault on the availability, integrity and confidentiality of the system [7].

[6]IDSs systems can be defined as attempts to identify intrusions, defined as unauthorized uses, misuses, or abuses of computer systems by either authorized or unauthorized users. Some IDSs monitor a single computer, while others monitor a particular network or many networks that form a Wide Area Network. IDSs detect intrusions by analyzing information about processes running in the computer as well as network traffic going in and out of the protected network. This type of information resides in audit trails, system tables, and network traffic summary logs. Intrusion detection is based on the assumption that the behavior of intruders differ from a legal user.

This paper defines for his study three main types for cyber analytics with support IDSs:misuse-based, Anomaly-based and Hybrid,then for each ML/DM method described we look for the different IDSs.

Machine Learning is defined as the study of algorithms that improve their performance with experience and are meant to computerize exercises ;the machine takes every necessary step consummately futhermore in a

maintained way[1]. As the definition says a machine learning learn by herself and with experience adapt herself for recognition for different attacks in an intrusion detection system. It is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed[2].
Data mining is identified as a solution to handling the analysis of data due to its adaptability and validity and it is now used extensively used for network security purposes [3].Data mining is employed into an intrusion detection system as a method of extracting the huge volumes of data that existing network traffic for further analysis[4]. As an application of machine learning, data mining holds a very significant position in intrusion detection, presenting methods of predicting future patterns based on past experiences[5]. So this paper describe the different ML/DM methods for cyber security which is the set of technologies and processes designed to protect computers from any type of attack or non functioning, and also the the type of datasets of network sources and systems are look for .

1.2.**Background**

DM(data mining) is a particular step in KDD(knowledge discovery of databases)-the application of specific algorithms for extracting patterns from data.
The pioneer of ML, Arthur Samuel, defined ML as a “field of study that gives computers the ability to learn without being explicitly programmed.” ML focuses on classification and prediction, based on known properties previously learned from the training data. ML algorithms need a goal (problem formulation) from the domain (e.g., dependent variable to predict). In this paper, it’s established that there should be three main phases for most ML method which are : training,validation and testing phases. The Predictive Model Markup Language (PMML) is developed and proposed by Data Mining Group to help predictive model sharing [8].It is based on XML and currently supports logistic regression and feed-forward neural network (NN) classifiers. There is three main types of ML/DM approaches : supervised,semi-supervised and unsupervised. the first one is when the label has a relation with the collected data,find the function that explain the data ; the second one is when the main task is to find patterns in unlabeled data,and the last is when a portion of data is labeled during the acquisition of data or by human experts. The following

datasets are studied:packet-level data, netflow data, and public data sets . Many ML/DM methods for cyber are studied like ANN(artificial neural networks), association rules and fuzzy association rules, Bayesian networks, clustering, decision trees, ensemble learning, evolutionary computation,
hidden markov models, inductive learning, naive bayes, sequential pattern mining, support vector machine(svm).

1.3.**Problem and Research questions**

The use of machine learning(ML) and data mining(DM) methods for cyber security intrusion detection demands to look at the type of cyber data(packet-level data, public data set, and netflow data), we have to take in account the complexity of algorithms of ML and DM methods and the performance of IDSs. In front of all these parameters, there is not yet a ML/DM method which is established as the most effective for intrusion detection.
For to find a start of solutions for this problem some questions have to be responded, like the following:
-Among these methods which is the most effective for the IDSs?
-what are the main factors to take in account for to realize an optimal IDS adapted ML/DM methods?
-which type of data sets are most effective or adaptive for what type of intrusion detection ?
-what are the most important factors related to the IDSs performance?
-what is the main criterion by which the ML/DM methods for cyber could be compared ?
-what makes it’s harder to use ML/DM methods in cyber domain?
-Among the ML/DM methods, what is the one on which the anomaly detector is based? And also the one on which the misuse detection is based?

2.RELATED WORKS

References	IDSs	Anomaly detection	IP Flows	Network Flow	Computational intelligence methods	Wireless network	Data mining (DM)
[1]	yes						
[2],[4]	yes	yes					
[3]			yes				
[5]			yes	yes			
[6]	yes	yes			yes		
[7]						yes	
[9]							yes
[28]	yes	yes					
[62]	yes						

This survey	yes	yes	yes	yes	yes	yes	Yes

References	Pcap data	Data sets	KDD data set	ANN (artificial neural networks)	Anomaly-detection	DBSCAN clustering	Association rules
[14],[15],[17], [16]	yes						
[18],[19]	yes	yes					
[20]			yes				
[21]			yes				
[22]				yes			
[27]					yes		
[46]						yes	
[32]							yes

[36]		yes					
[35]							yes
[51]						yes	
This survey	yes	yes	yes	yes	yes	yes	yes

3.METHODOLOGY EMPLOYED

The authors of this article are presenting the ML/DM methods for cyber security, in particular for intrusion detection. To make this study they begin at the introduction to define the cyber security, and they show the different types of cyber analytics such as misuse-based,anomaly based and hybrid. After this, they show the main addition brought by this survey and compare it with prior surveys(some related already existing and differences). Then, the others parts(called here sections) of the survey are announced :here, we have in section 2 the major steps in ML/DM, then we have in section 3 cyber security data sets for ML and DM, then section 4 present ML and DM Methods for cyber, computational complexity of ML/DM methods, then it's about some recommendations and observations and finally we have conclusion.

The section 2 presents the evolution of ML/DM methods, and the major steps are described,an emphasis is on KDD processes for DM(data mining) and on the phases of ML(machine learning);it's shown the three main types of ML/DM approaches:supervised learning, unsupervised learning and semi-supervised learning.

Then in section 3 the different of data sets used when we use ML/DM methods are described:packet-level data, public data sets, and network flow data.

Section 4 is the main part of this article which presents the ML and DM methods for cyber :artificial neural networks,association rules and fuzzy association rules , bayesian network, clustering, decision trees, ensemble learning, evolutionary computation, hidden markov models, naive bayes, sequential pattern mining, svm(support vector machine)
section 5 presents the computational complexity of ML/DM methods,here when training prediction models we have some main factors :time complexity, incremental update capability, and generalization capacity.
Section 6 gives the observations and recommendations about all what have been told and described previously, some observations are done related to data set, related to IDSs performance, and also comparison criteria are made. Finally some peculiarities are raised on ML and DM methods for cyber.
The last section is the conclusion which resumes the entire work.

4.RESULTS FOUND AND PERSPECTIVES

Some results have been found about this survey paper, considering the IDSs performance, there are two main factors to take in account :

the first is the type and level of input data . In this survey, some related works used DARPA and KDD data sets because they are easy to obtain and contain network level data(either tcpdump or netflow) as well as OS-level data, as cited in this survey:"As a result, it is advantageous that an IDS be able to reach network-level and kernel-level data. If only NetFlow (much easier to obtain and process) data are available for the IDS, these data must be augmented by network-level data such as network sensors that generate additional features of packets or streams. As perspective If possible, the network data should be augmented by OS kernel-level data. As discovered, several studies approached the intrusion detection problem by examining OS-level commands (i.e., host-based IDS), not network packets."

the second main factor is the type of ML and DM algorithms employed and the overall system design.

As comparison criteria, we have several criteria by which the ML/DM methods for cyber could be compared: Accuracy, time for training model, time for classifyng an unknown instance with a trained model, understandability of the final solution(classification)

The time for training a model is an important factor due to ever changing cyber-attack types and features. Even anomaly detectors need to be trained frequently, perhaps incrementally, with fresh malware signature updates.
Time for classifying a new instance is an important factor that reflects the reaction time and the packet processing power of the intrusion detection system. Understandability or readability of the classification model
is a means to help the administrators examine the model features easily in order to patch their systems more quickly

About **Peculiarities of ML and DM for Cyber The cyber** domain has some peculiarities that make those methods harder to use. Those peculiarities are especially related to how often the model needs to be retrained and the availability of labeled data.
“In most ML and DM applications, a model (e.g., classifier) is trained and then used for a long time, without any changes. In those applications, the processes are assumed to be quasi stationary and therefore the retraining of the model does not happen often.” from this survey in section .
There are many domains in which it is easy to obtain training data, and in those domains ML and DM methods usually thrive (e.g., recommendations that Amazon makes for its clients). In other areas where data are difficult to obtain (e.g., health monitoring data for machinery or aircraft), applications of ML and DM can be abundant. In the cyber domain, much data could be harvested by putting sensors on networks (e.g., to get NetFlow or TCP), which although not an easy task is definitely worthwhile. Often, an anomaly detector is based on a clustering method.
Among clustering algorithms, density-based methods (e.g., DBSCAN) are the most versatile, easy to implement, less parameter or distribution dependent, and have high processing speeds. In anomaly detectors, one-class SVMs also perform well and much can be learned by extracting association rules or sequential patterns from available normal traffic data.

Among misuse detectors, because the signatures need to be captured, it is important that the classifier be able to generate readable signatures, such as branch features in a decision tree, genes in a genetic algorithm, rules in Association Rule Mining,

The designers of the system should investigate whether the training data are of good enough quality and have statistical properties that can be exploited (e.g., Gaussian distribution). It is also important to know whether the required system will work on-line or off-line. Answers to such questions will determine the most suitable ML approach. In the opinion of the authors of this paper, the network data cannot be properly modeled using a simple distribution (e.g., Gaussian) due to the fact that, in practice, a single network packet might contain a payload that can be associated to dozens of network protocols and user behaviors

5.MOST IMPORTANT PUBLICATIONS FOR THIS WORK

[89] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. San Mateo, CA, USA: Morgan Kaufmann,2011. 6times

[52] 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. 5 times

[62] 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. 5 times

[92] R. Agrawal and R. Srikant, “Mining sequential patterns,” in Proc. IEEE 11th Int. Conf. Data Eng., 1995, pp. 3–14. 4 times

[43] 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. 3times

[87] W. Lee, S. Stolfo, and K. Mok, “A data mining framework for building intrusion detection models,” in Proc. IEEE Symp. Secur. Privacy, 1999, pp. 120–132. 3 times

[91] 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. 3 times

[42] 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. 3 times

[56] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, “EXPOSURE: Finding malicious domains using passive DNS analysis,” presented at the 18th Annu. Netw. Distrib. Syst. Secur. Conf., 2011. 3 times

[35] D. Apiletti, E. Baralis, T. Cerquitelli, and V. D’Elia, “Characterizing network traffic by means of the NetMine framework,” Comput. Netw., vol. 53, no. 6, pp. 774–789, Apr. 2009. 3 times

[16] Snort 2.0. Sourcefire [Online]. Available: http://www.sourcefire.com/technology/whitepapers.htm, accessed on Jun. 2014. 3 times

References used for this report :

[1] Machine Learning Techniques for Intrusion Detection: A Comparative Analysis
Yasir Hamid, M Sugumaran, Ludovic Journaux
HAL Id: hal-01392098
https://hal.archives-ouvertes.fr/hal-01392098

[2] « Machine learning -Wikipedia, the free encyclopedia. » [Online]. Available: https://en.wikipedia.org/wiki/Machine_learning.

[3]R.J. Manish, and H.T. Hadi, “A review of network traffic analysis and prediction techniques” unpublished.

[4]S. Choudhury and A.Bhowal
,
“Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network intrusion Detection.”In: Smart Technologies and Management for Computing,
Communication, Controls, Energy and Materials, IEEE,2015, pp. 89-95.

[5]S.B. Kotsiantis, I.D. Zaharakis and P .E. Pintelas, “
Machine Learning: a Review of Classification and combining Techniques,”Artificial Intelligence Review ‘vol. 26’, November 2006, pp. 159-190.

[6]A Survey on Machine Learning Techniques for
Intrusion Detection
Systems
Jayveer Singh[1],Manisha J.Nene[2] Department of Computer Engineering, DIAT, Pune, India, 411025 [1, 2]

[7]R. Heady, G. F. Luger, A. Maccabe, and M. Servilla, The architecture of a network level intrusion detection system. Department of Computer Science, College of Engineering, University of New Mexico, 1990.

[8] A. Guazzelli, M. Zeller, W. Chen, and G. Williams, “PMML an open standard for sharing models,” R J., vol. 1, no. 1, pp. 60–65, May 2009.è-