Report of the subject 6 : A survey of Data Mining and Machine Learning Methods for cyber Security Intrusion Detection.

1 .Introduction

context :

this survey is taken from the IEEE (Institute of Electrical and Electronics Engineers) COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 2, SECOND QUARTER 2016.

The massive increase in the rate of novel cyber attacks has made data-mining-based techniques a critical component in detecting security threats[114].

Anna L.Buczak, and Erhan Guven who are the members the IEEE organisation (Institute of Electrical and Electronics Engineers) , describe a focused litterature survey of machine learning(ML) and Data Mining(DM) methods for cyber analytics in support of intrusion detection .Then, this survey defines the differents types of cyber analytics intrusion detection(misuse,anomaly and hybrid),relate the differents methods used in machine leaning and data mining and analyse for each method the actions of differents types of detection(misuse and anomaly/hybrid).

This document is intended for readers who wish to begin research in the field of ML/DM for cyber intrusion detection. As such, great emphasis is placed on a thorough description of the ML/DM methods, and references to seminal works for each ML and DM method are provided.

background

here we have to see small evolution of how increase DM and ML for cyber attacks intrusion detection.

some confusions exist between ML,DM and KDD(knowledge discovery in databases) but let's show a definition of these terms :The pioneer of ML, Arthur Samuel, defined ML as a "field of studythat 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.

Another definition of ML from according to IOANNIS KAVAKIOTIS[115] Machine learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term "machine learning" is identical to the term "artificial intelligence", given that the possibility of learning is the main characteristic of an entity called intelligent in the broadest sense of the word. The purpose of machine learning is the construction of computer systems that can adapt and learn from their experience .

DM focuses on the discovery of previously unknown properties in the data.

It does not need a specific goal from the domain, but instead focuses on finding new and interesting knowledge

now that these main terms are defined let's say about their classification:some people can view ML as the older sibling of DM because the term ML (machine learning) has

been in use since 1960s and the term DM(data mining) was introduced in late 1980s,and thirdly the first conference of KDD took place in 1989. then DM is more popular and recent than ML,it's why some researchers label their work DM rather than ML ,it could also be the reason that when queries "machine learning" AND cyber and "data min-
ing" AND cyber were performed on Google Scholar, the first
retrieved 21,300 results and the second retrieved 40,800 results.
Because papers retrieved by the two queries are not significantly different, the studied methods here are the ML/DM methods.
some surveys about the machine learning and data mining have been published and they are cited in this survey,but this last comes with some particularities,then we can cite some main apports of this survey like that :
firstly, this survey has the recent informations about the machine learning and data mining methods for intrusion detection because of its recents edition and publication .this paper concentrates only on machine learning and data mining methods,in addition to the anomaly detection ,signature-based and hybrid methods are depicted.
Also this paper present the methods that work on any type of cyber data;our paper also includes the methods for recognition of type of the attack(misuse) and for detection of an attack(intrusion), our paper presents the full and latest list of ML/DM methods that are applied to cyber security.the machine learning and data mining methods in this paper are fully applicable to the intrusion and misuse detection problems in both wired and wireless networks

problem and research questions :
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.
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. If only NetFlow data are available, these data must be augmented by network-level data such as network sensors that generate additional features.
Some research questions :
-which parameters to take in accounts for to establish a most-effctive method for cyber security intrusion detection in cyber ?
-because the machine learning and data mining can not work whithout data,how to establish a better representation of data for to optimize the time for obtaining of these data sets?
-how to better increase the network flow data by network-level data and by OS kernel-level data to optimize the intrusion detection ?
-How to investigate and find the better method for a fast incremental learning that could be used for daily updates of models for misuse and anomaly detection ?

2.Related works

References	Cyber analytics	ML and DM methods	datas(type)	Network topology	Year of publication
[2]	yes			yes	
[3]		Yes, in internet traffic			Used surveys(2004-2007),published in
[4]	Yes,anomaly-based network intrusion	Yes,small,approaches			
[5]	yes		Yes,network flow data		
[6]	yes	yes		yes	
[7]				yes	
[1]	Yes				
This paper	yes	yes	Cyber data and interrnet protocol flows		2016

Methodology employed

-Firstly we have to put at the top of survey the refence of the survey,like following : topic of source,number of volume,number of article in this source,and edition's house and the publication's year.
-then let's put the title of the survey and beside the authors about this survey.

-then we have abstract : it resumes what contains the survey and the aspects with great emphasis.
-after this, the key words about the survey are cited.

Then, we have the introduction and the rest of the survey like following :

I-Introduction
the situation of the survey is given,and the context is described ;
after this , we have the definitions of key words about survey
-a comparison between the survey and others related works is done ,and an emphasis is also mentioned about specifity of survey.
-a quiet description of the date accept , date publication and date of current version about the survey.the different partners and laboratories associated to the survey are mentioned,email address available.
-the next thing to do is to give the plan or the different sections of the remain of the survey.
-the author has to prevent about the rights of use of the survey,tell is open source or limited.tell in which cases the contain maybe used.

II-section 2 title considering that introduction was the first section
speak about what question is…
here,the evolution or background about the survey topic,all about the main steps passed for the increasing of the studied domain until this survey's apparition.
III-section 3 title
firstly, we have a small introduction describing how the differents authors applied the differents methods.
in this part, it's about of the deep of the studied domain with all the elements or omponents depending of the subject .every component is a subtitle of the section and any subtitle is more detailed in his text,then we can have a presentation like that :
-compenent 1
…
-component2
...
(…)
etc
IV-section 4 title
a little introduction about of differents elements is done, and it's shown some papers which have worked on it. the references are associated to each component.

A great emphasis is mentioned about the main elements concerning the survey
also here each subtitle define a component and speaks about it ,and describe more and more.

V-section 5 title

this section describe the complexity meet in the diffenrents components studied above.

VI-section 6 title

some illustrations are seen in the table with some comparisons.
it’s about the observations and recommandations, the extent of this part is about to conclude for what what component or method is better to use or is more effective for studied. In each subtitle,it’s associated observations related.

VII-section 7 title

it’s about conlusion , here all is related,the results about the research,the encountered problems and research questions related to the problem.

References :

here it’s about of all the sources of documents used and the names of their authors for the redaction.

4.Results found and perspectives

Results found :

the extent of papers found on ML and DM for cyber intrusion detection shows that these methods are a prevalent and growing research area for cyber security.

Netflow does not have such a rich set of features as tcpdump,DARPA or KDD data nor the features necessary for detecting particular signatures in misuse detection .(its features are limited to flow information generated by higher-end routers)

About factors related to IDS performance :

-One of the most important factors related to the performance of IDSs is the type and level of the input data

As another result, it is advantageous that an IDS(intrusion detection system) be able to reach network- and kernel – level data.

There is another result :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. 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.

About comparison criteria :

There are several criteria by which the ML/DM methods for cyber could be compared:

• Accuracy
• Time for training a model
• Time for classifying 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.

For 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.

A fertile area of research would be to investigate the methods of fast incremental learning that could be used for daily updates of models for misuse and anomaly detection.
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.However there is a problem with the sheer volume of that data—there is too much data to store (terabytes per day).Another problem is that some of the data need to be labeled to be useful, which might be a laborious task.

Data for training definitely need to be labeled, and this is even true for pure anomaly detection methods. These methods have to use data that are normal; they cannot develop a model with the attack data intermingled. Additionally, because they have to be tested with novel attacks, some novel attack data are required as well.

Otherwise, the best possible available data set right now is the KDD 1999 corrected data set.
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 online or offline.
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.

Most used 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. 6 times

[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