Applications of Machine Learning and Data Mining for Cyber Security

Ruby Dahiya*
Anamika**

Abstract

Security is an essential objective in any digital communication. Nowadays, there is enormous information, lots of protocols, too many layers and applications, and massive use of these applications for various tasks. With this wealth of information, there is also too little information about what is important for detecting attacks. Methods of machine learning and data mining can help to build better detectors from massive amounts of complex data. Such methods can also help to discover the information required to build more secure systems, free of attacks. This paper will highlight the applications of machine learning and data mining techniques for securing data in huge network of computers. This paper will also present the review of applications of data mining and machine learning in the field of computer security. The papers which will be reviewed here, present the results of various techniques of data mining and machine learning on different performance parameters.

Keywords: Data mining, Machine Learning, Artificial Neural Networks, Classification, Clustering, Inductive Learning, Evolution Learning, Support Vector Machine.

I. Introduction

As technology moves forward user become more technical aware then before. People communicate and corporate efficiently through the internet using their PC's, PDs or mobile phones. Through these digital devices link by the internet, hacker also attack personal privacy using a variety of weapons such as virus, worms, botnet attacks, spam and social engineering platforms. These forms of attack can be categorized into three groups- Stilling confidential information, manipulating the components of cyber infrastructures and denying the functions of infrastructure. There are three approaches to deal with these attacks: signature-based, anomaly-based and hybrid. The signature based detection system use the particular signature of an attack, hence are unable to detect unknown attacks. The anomaly-based system detects the anomalies as the deviation from the normal behavior so they can detect unknown attacks as well. The main disadvantage of these systems is high false alarm rates (FAR). The hybrid approach uses the combination of both signature-based and anomaly-based techniques. These types of system have high detection rate of known attacks and low false positive rates for unknown attacks. The literature review shows that most of the techniques were actually hybrid. The security mechanisms are also categorized as: network based and host based. A network-based system monitors the traffic through the network devices. A host based system monitors the processes and the file related activities associated with a specific host. However building a defense system for discovered attacks is not easy because of constantly evolving cyber attacks. The figure 1 depicts the cyber security mechanism.

This paper is intended for readers who wish to begin research in the field of machine learning and data mining for cyber security. This paper highlights ML and DM techniques used for cyber security. The paper describes ML and DM techniques in reference to anomaly method and signature based hybrid methods however the in depth description of these methods is in the paper of Bhuyan et al. [1]. This paper focuses on cyber intrusion detection for both wired and wireless networks. The paper Zhang el al. [2] focuses more on dynamic networking.

Ruby Dahiya*
Associate Professor (IT)
Institute of Information Technology & Management

Anamika**
Assistant Professor (IT)
Institute of Information Technology & Management

Figure1. Cyber Security System

The paper is organized as follow: section II highlights the procedure of Machine Learning and Data Mining. Section III describes the techniques of ML and DM. Section IV presents and discusses the comparative analysis of individual technique and related work. Section V presents the conclusion.

II. Machine Learning and Data mining Procedure

The ML and DM are two terms that are often confused because generally, they both have same techniques. Machine Learning, a branch of artificial intelligence, was originally employed to develop hniques to enable computers to learn. Arthur Samuel in 1959 defined Machine Learning as a "field of study that gives computers the ability to learn without being explicitly programmed"[3]. ML algorithm applies classification followed by prediction, based on known properties learned from the training data. ML algorithms need a well defined problem from the domain where as DM focuses on the unknown properties in the data discovered priory. DM focuses on finding new and interesting knowledge. An ML approach consists of two phases: training and testing. These phases include classification of training data, feature selection, training of the model and use of model for testing unknown data.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The following are areas in which data mining technology may be applied or further developed for intrusion detection

- Development of data mining algorithms for intrusion detection: Data mining algorithms can be used for misuse detection and anomaly detection. The techniques must be efficient and scalable, and capable of handling network data of high volume, dimensionality and heterogeneity.
- Association and correlation analysis and aggregation to help select and build discriminating attributes: Association and correlation mining can be applied to find relationships between system attributes describing the network data. Such information can provide insight regarding the selection of useful attributes for intrusion detection.
- Analysis of stream data: Due to the transient and dynamic nature of intrusions and malicious attacks, it is crucial to perform intrusion detection in the data stream environment. It is necessary to study what sequences of events are frequently encountered together, finding sequential patterns, and identify outliers.
- Distributed data mining: Intrusions can be launched from several different locations and targeted to many different destinations. Distributed data mining methods may be used to analyze network data from several network locations in order to detect these distributed attacks.

- Visualization and querying tools: Visualization tools should be available for viewing any anomalous patterns detected. Intrusion detection systems should also have a graphical user interface that allows security analysts to pose queries regarding the network data or intrusion detection results.

III. Techniques of ML and DM

This section focuses on the various ML/DM techniques for cyber security. Here, each technique is elaborated with references to the seminal work. Few papers of each technique related to their applications to cyber security.

*A. Artificial Neural Networks:*Neural Networks follow predictive model which are based on biological modeling capability and predicts data by a learning process. The Artificial Neural Networks (ANN) is composed of connected artificial neurons capable of certain computations on their inputs [4]. When ANN is used as classifiers, the each layer passes its output as an input to the next layer and the output of the last layer generates the final classification category.

ANN are widely accepted classifiers that are based on perceptron [5] but suffer from local minima and lengthy learning process. This technique of ANN is used for as multi-category classifier for signature-based detection by Cannady [6]. He detected 3000 simulated attacks from a dataset of events. The findings of the paper reported almost 93% accuracy and error rate 0.070 root mean square. This technique is also used by Lippmann and Cunningham [27] for anomaly detection. They used keyword selection based on statistics and fed it to ANN which provides posterior probability of attack as output. This approach showed 80% detection rate and hardly one false alarm per day. Also, a five-stage approach for intrusion detection was proposed by Biven et al. [8] that fully detected the normal behavior but FAR is 76% only for some attacks.

B. Association Rules and Fuzzy Association Rules: Association Rule Mining was introduced by Agarwal et.al. [9], as a way to find interesting co-occurrences in super market data to find frequent set of items which bought together. The traditional association rule mining works only on binary data i.e. an item was either present in the transaction will be represented by 1 or 0 if not. But, in the real world applications, data are either quantitative or categorical for which Boolean rules are unsatisfactory. To overcome this limitation, Fuzzy Association Rule Mining was introduced [10], which can process numerical and categorical variables.

An algorithm based on Signature Apriori method was proposed by Zhengbing et al. [11] that can be applied to any signature based systems for the inclusion of new signatures. The work of Brahmi [12] using multidimensional Association rule mining is also very promising for creating signatures for the attacks. It showed the detection rate of attacks types DOS, Probe, U2R and R2L as 99%, 95%, 75% and 87% respectively. Association rule mining is used in NETMINE [35] for anomaly detection. It applied generalization association rule extraction based on Genio algorithm for the identification of recurring items. The fuzzy association rule mining is used by Tajbakhsh et al. [38] to find the related patterns in KDD 1999 dataset. The result showed good performance with 100 percent accuracy and false positive rate of 13%. But, the accuracy falls drastically with fall of FPR.

C. Bayesian Networks: A Bayesian is a graphical model based on probabilities which represents the variables and their relationships [15], [16]. The network is designed with nodes as the continuous or discrete variables and the relationship between them is represented by the edges, establishing a directed acyclic graph. Each node holds the states of the random variable and the conditional probability form.

Livadas et al. [17] presented comparative results of various approaches to DOS attack. The anomaly detection approach is mainly reactive whereas signature-based is proactive. They tried to detect botnets in Internet Relay Chat (IRC) traffic data. The analysis reported the performance of Bayesian networks as 93% precision and very low FP rate of 1.39%.Another IDS based on Bayesian networks classifiers was proposed by Jemili et al. [18] with performances of 89%, 99%, 21% and 7% for DOS, Probe, U2R and R2L respectively. Benferhat [19] also used this approach to build IDS for DOS attack.

D. Clustering: Clustering is unsupervised technique to find patterns in high-dimensional unlabeled data. It is used to group data items into clusters based on a similarity measure which are not predefined.

This technique was applied by Blowers and Williams [20] to detect anomaly in KDD dataset at packet level. They used DBSCAN clustering technique. The study highlighted various machine learning techniques for cyber security. Sequeira and Zaki [21] performed detection over shell commands data to identify whether the user is a legitimate one or intruder. Out of various approaches for sequence matching, the longest common sequence was the most appropriate one. They stated the performance in terms of 80% accuracies and 15% false alarm rate.

E. Decision Trees: It is a tree like structure where the leaf node represents or predicts the decision and the non-leaf node represents the various possible conditions that can occur. The decision tree technique has simple implementation, high accuracy and intuitive knowledge expression. This expression is large for small trees and less for deeper and wider trees. The common algorithms for creating decision tree are ID3 [22] and C4.5 [23].

Kruegel and Toth [24] proposed clustering along with decision tree approach to build a signature detection system and compared its performance to SNORT2.0.The speed up varies from 105% to 5 %, depending on the traffic. This paper showed that the combination of decision trees with clustering technique can prove an efficient IDS approach. The decision tree approach using WEKA J48 program was also used in EXPOSURE [25] to detect the malicious domains like botnet command, scam hosts, phishing sites etc. Its performance is satisfactory in terms of accuracy and FAR.

F. Ensemble Learning: It is a supervised machine learning paradigm where multiple learners are trained to solve the same problem. As compared with ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses and combine them to use.

An outlier detector was designed to classify data as anomaly as well as to classify it to one of the attack labels of KDD dataset by Zhang et al. [26] with the use of Random Forests. The Random forest was used as the proximity measure. The accuracy for the DOS, Probe, U2R and R2L attacks were 95%, 93%, 90% and 87% respectively. The FAR is 1%.

G. Evolutionary Computation: It is the collective name for a range of problem-solving techniques like Genetic Algorithms, genetic programming, particle swarm optimization, ant colony optimization and evolution strategies based on principles of biological evolution.

The signature-based model was developed by Li [27] with genetic algorithms used for evolving rules. Abraham et al. [28] also used genetic programming techniques to classify attacks in DARPA 1998 intrusion detection dataset.

H. Inductive Learning: It is a learning method where learner starts with specific observations and measures, begins to detect patterns and regularities, formulates some tentative hypothesis to be explored and ends up with development of some general conclusion and theories. Inductive learning moves from bottom-up that is from specific observations to broader generalizations and theories. Repeated Incremental Pruning to Produce Error Reduction RIPPER [29] applies separate and conquer approach to induce rules in two-class problems. Lee et al. [31] provided a framework for signature-based model using various machine learning and data mining techniques like inductive learning, association rules, sequential pattern mining etc.

I. Naïve Bayes: It is a simple probabilistic classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.Panda and Patra [31] presented the comparison of Naïve Bayes with NN classifier and stated that Naïve Bayes performed better in terms of accuracy but not in FAR. Amor et. al. [32] used Bayesian network as naïve bayes classifier. The paper stated accuracy of 98% with less than 3% false alarm rate.

J. Support Vector Machine: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. *In other* words, given

Table1. Analysis of ML and DM techniques

ML/DM Technique	Method	Data Set	Evaluation Metric	Work
ANN	Signature based	Network Packet level	Acc., RMS	Cannady
ANN	Anomaly	DARPA 1998	DR, FAR	Lippmann & Cunningham
ANN	Anomaly	DARPA 1999	DR, FAR	Bivens et. al.
Association Rules	Signature based	DARPA 1998	DR	Brahmi et. al.
Association Rules	Signature based	Signature attacks	Runtime	Zhengbing et. al.
Association Rules - Fuzzy	Hybrid	KDD 1999 (corrected)	Acc., FAR	Tajbakhsh et. al.
Bayesian Network	Signature based	Tcpdump- botnet traffic	Precision, FAR	Livadas et. al.
Bayesian Network	Signature based	KDD 1999	DR	Jemili et. al.
Clustering- density based	Anomaly	KDD 1999	DR but no actual FAR	Blowers and Williams
Clustering – Sequence	Anomaly	Shell Commands	Acc., FAR	Sequeira and Zaki
Decision Tree	Signature based	DARPA 1999	Speedup	Kruegel and Toth
Ensemble – Random Forest	Hybrid	KDD 1999	Acc., FAR	Zhang et. al.
Evolutionary Computing (GA)	Signature based	DARPA 2000	Acc.	Li
Evolutionary Computing (GP)	Signature based	DARPA 1998	FAR	Abraham et. al.
Inductive Learning	Signature based	DARPA 1998	Acc.	Lee et. al.
Naïve Bayes	Signature based	KDD 1999	Acc., FAR	Panda & Patra
Naïve Bayes	Anomaly	KDD 1999	Acc., FAR	Amor et. al.
Support Vector Machine	Signature based	KDD 1999	Acc.	Li et. al.
Support Vector Machine	Anomaly	DARPA 1998	Acc., FAR	Hu et. al.

labeled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples.

An SVM classifier was built to classify KDD 1999 dataset by Li et. al.[33] using ant colony optimization for the trainee. This study showed 98% accuracy, however it is not performing well for U2R attacks. RSVM(Robust Support Vector Machine) was used as anomaly classifier by Hu et. al.[34] which showed a better performance with noise having 75% accuracy with no false alarms.

IV. Comparative Analysis And Discussion

The analysis of the work using of ML and DM for cyber security highlights few facts about the growing research area in this field. From the comparative analysis presented in Table 1, it is obvious that the DARPA 1998, DARPA 1999, DARPA2000 KDD 1998, KDD 1999 are the favorite choices of most of the researchers for the dataset for IDS. Most of the researches have used accuracy, detection rate, false alarm rate as the evaluation criteria. There have been multiple approaches that are applied for both anomaly and signature-based detection. Several approaches are appropriate for signature-based others are for anomaly detection. But, the answer to the question about determination of most appropriate approach depends on multiple factors like the quality of the training data, properties of that data, working of the system(online or offline) etc.

V. Conclusions

In this paper, we survey a wide spectrum of existing studies on machine learning and data mining techniques applied for the cyber security. Based on this analysis we then outline key factors that need to be considered while choosing the technique to develop an IDS. These are the quality and properties of the training data, the system type for which the IDS has to be devised and the working nature and environment

of the system. There is a strong need to develop strong representative dataset augmented by network data level. There is also a need to regular updating of the models for the cyber detection using some fast incremental learning ways.

References

1. M. Bhuyan, D. Bhattacharyya, and J. Kalita, "Network anomaly detection:Methods, systems and tools," *IEEE Commun. Surv. Tuts.*, vol. 16, no. 1, pp. 303–336, First Quart. 2014.
2. Y. Zhang, L. Wenke, and Y.-A. Huang, "Intrusion detection techniques for mobile wireless networks," *Wireless Netw.*, vol. 9, no. 5, pp. 545–556, 2003.
3. J. McCarthy, "Arthur Samuel: Pioneer in Machine Learning," AI Magazine, vol. 11, no. 3, pp. 10-11, 1990.
4. K. Hornik,M. Stinchcombe, and H.White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, pp. 359–366,1989.
5. F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6,pp. 386–408, 1958.
6. J. Cannady, "Artificial neural networks for misuse detection," in *Proc 1998 Nat. Inf. Syst. Secur. Conf.*, Arlington, VA, USA, 1998, pp. 443–456.
7. 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.
8. 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.
9. R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. Int. Conf. Manage. Data Assoc. Comput. Mach. (ACM)*, 1993, pp. 207–216.
10. C. M. Kuok, A. Fu, and M. H. Wong, "Mining fuzzy association rules in databases," *ACM SIGMOD Rec.*, vol. 27, no. 1, pp. 41–46, 1998.
11. 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.
12. 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.
13. 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.
14. A. Tajbakhsh, M. Rahmati, and A. Mirzaei, "Intrusion detection using fuzzy association rules," *Appl. Soft Comput.*, vol. 9, pp. 462–469, 2009.
15. D. Heckerman, *A Tutorial on Learning with Bayesian Networks*. New York, NY, USA: Springer, 1998.
16. F. V. Jensen, *Bayesian Networks and Decision Graphs*. New York, NY, USA: Springer, 2001.
17. C. Livadas, R.Walsh, D. Lapsley, andW. Strayer, "Usingmachine learning techniques to identify botnet traffic," in *Proc 31st IEEE Conf. Local Comput. Netw.*, 2006, pp. 967–974.
18. 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.

19. S. Benferhat, T. Kenaza, and A. Mokhtari, "A Naïve Bayes approach for detecting coordinated attacks," in *Proc. 32nd Annu. IEEE Int. Comput. Software Appl. Conf.*, 2008, pp. 704–709.

20. M. Blowers and J. Williams, "Machine learning applied to cyber operations," in *Network Science and Cybersecurity*. New York, NY, USA: Springer, 2014, pp. 55–175.

21. 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.

22. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

23. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

24. C. Kruegel and T. Toth, "Using decision trees to improve signature based intrusion detection," in *Proc. 6th Int. Workshop Recent Adv. Intrusion Detect.*, West Lafayette, IN, USA, 2003, pp. 173–191.

25. 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.

26. 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.

27. W. Li, "Using genetic algorithms for network intrusion detection," in *Proc. U.S. Dept. Energy Cyber Secur. Group 2004 Train. Conf.*, 2004, pp. 1–8.

28. A. Abraham, C. Grosan, and C. Martin-Vide, "Evolutionary design of intrusion detection programs," *Int. J. Netw. Secur.*, vol. 4, no. 3, pp. 328–339, 2007.

29. W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, Lake Tahoe, CA, USA, 1995, pp. 115–123.

30. 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.

31. M. Panda and M. R. Patra, "Network intrusion detection using Naïve Bayes," *Int. J. Comput. Sci. Netw. Secur.*, vol. 7, no. 12, pp. 258–263, 2007.

32. N. B. Amor, S. Benferhat, and Z. Elouedi, "Naïve Bayes vs. decision trees in intrusion detection systems," in *Proc ACMSymp. Appl. Comput.*, 2004, pp. 420–424.

33. 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.

34. 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.