International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

A Survey : Data Mining and Machine Learning Methods for Cyber Security

Ashish Prajapati[1], Shital Gupta[2]

[1]M.Tech Scholar, School of Research & Technology People's University, Bhopal, Madhya Pradesh, India
[2]Assistant Professor, School of Research & Technology People's University, Bhopal, Madhya Pradesh, India

Article Info
Volume 7, Issue 2
Page Number: 24-34
Publication Issue :
March-April-2021
Article History
Accepted : 01 March 2021
Published : 08 March 2021

ABSTRACT

This survey paper describes the literature survey for cyber analytics in support of intrusion detection of machine learnings (ML) and data mining (DM) methods. Short ML/DM method tutorial details will be given. Documents representing each method were categorized, read and summarized based on the number of citations and significance of an evolving method. Since data is so important.

Keywords: Machine Learnings, Data Mining, Cyber Security, . Novel (Zero-Day) Attacks

I. INTRODUCTION

This report shows the findings of a literature review of cyber security applications in machine learning (ML) and data mining (DM) methods. The ML/DM methods and multiple applications for each method are listed for the detection of cyber intrusion issues. The complexity of various ML/DM algorithms is discussed and the paper offers a set of comparative parameters for the ML/DM methodology and a set of guidelines on the best possible methods for solving cyber problems.

Cyber security is the collection of technologies and processes designed to protect against attacks, unauthorized entry, modification or degradation of computers, networks, programs and data. Network Security Systems and device (host) security systems are integrated into cyber security systems. Each one at least has an intrusion detection system, firewall and antivirus applications (IDS). The IDSs help discover, classify and detect unauthorized information systems use, replication, modification and destruction [1]. The safety violations include external (external attacks) and internal intrusions (attacks from within the organization).

There are three principal forms of IDS support cyber analytics: maluses based (also known as signature-based), anomalous and hybrid-based. Misuse-based techniques are constructed by using signatures of these attacks to detect known attacks. They detect known attacks without creating an excessive number of false alarms. They are successful. Manual updates of the database with rules and signatures are also essential. Novel (zero-day) attacks cannot be detected by misuse techniques.

Anomaly-orientated approaches model normal network and system activity and classify abnormalities as abnormalities. Their ability to detect zero-day attacks makes them appealing. Another benefit is that typical activity profiles are adapted to each device, application or network and make it difficult for attackers to know what behaviours they can do without detection. In addition, the data used to alert anomaly-based techniques (novel attacks) for signatures for detectors may also be used. The key drawback of anomaly-based techniques is that there is a potential to have high false alarm rates (FARs) because the device activity which was previously unseen (but legitimate).

This study focuses on the ML and DM cybersecurity approaches with a focus on and examples of the ML/DM methods. A number of documents detailing these approaches, including many analyses, have been released. The emphasis of the paper is on publications that follow those standards, unlike previous reviews. Google Scholar queries using "machine learning" and "cyber," and "data mining" and cyber were conducted. Highly quoted papers were emphasized in particular because the techniques discussed were common. However, this focus may neglect major new and evolving techniques, so some of these papers have also been selected. Overall, documents were selected so that at least one and, preferably, a few representative papers were available for each ML/DM category mentioned later.

II. METHODS AND MATERIAL

Cybersecurity Attack:

In the early days, malware was written simply to demonstrate security vulnerabilities or, in certain cases, to demonstrate technical skills. The primary purpose of malware today is to steal confidential information from others for the benefit of individuals, finance or company [3][4]. For example, malware is sometimes used to target government or corporate websites for the collection or disruption of guarded information. Malware is used in other situations to collect personal data such as social security numbers or credit card numbers even against individuals. As broadband broad band Internet connections grow more cheaply and quickly, malware is increasingly built not only for information purposes but purely for profit [5]. For example, most common malware is designed for controlling black-market computers such as e-mail spam sending, web navigation behaviours and the display of unsolicited ads. In 2012, 26 million new malwares are identified on the basis of the Anti-Phishing Community report.

Spam refers to sending irrelevant, inappropriate and unsolicited messages to thousands or millions of recipients. Spam has turned out to be a highly profitable market since spam is sent anonymously with no costs involved beyond the management of mailing lists. Due to such low barrier to entry, spammers are numerous, and the volume of unsolicited mail has grown enormously. In the year 2011, the estimated figure for spam messages is around seven trillion [2]. This figure includes the cost involved in lost productivity and fraud, and extra capacity needed to cope with the spam. Today, most widely recognized form of spam is email spam. According to the Message Anti-Abuse Working Group report [1], between 88–92% of email messages sent in the first half of 2010 carried spam.

- Phishing is a way of attempting to acquire sensitive information such as username, password or credit card details by masquerading as a trustworthy entity. Most phishing scams rely on deceiving a user into visiting a malicious web site claiming to be from legitimate businesses and agencies. Unsuspecting user enters private information in the malicious web site which is then subsequently used by malicious criminals. Most methods of phishing use some form of technical deception designed to make a link in an email (and spoofed website) appear to belong to a legitimate organization, such as well-known bank. Misspelled URLs or the use of sub-domains are common tricks used by phishers. The Anti-Phishing technical report stated that, there was a visible trend of phishers in 2011 to hide their intentions by avoiding the use of obvious IP host to host their fake login pages. Instead, the phishers preferred to host on a compromised domain to avoid detection. It is reported that there was 16 percent drop in the number of phishing URLs containing the spoofed company name in the URL. These combined trends show how phishers are adapting as users becoming more informed and knowledgeable about the traits of a typical phish.

- Drive-by Downloads concerns the unintended downloads of malware from the Internet and have been increasingly used by the attackers to spread malware fast. Drive-by downloads happen in a variety of situations; for example, when a user visits a website, while viewing an email message by user or when users click on a deceptive pop-up window. However, the most popular drive-by downloads occur by far when visiting websites. An increasing number of web pages have been infected with various types of malware. According to Osterman Research survey [3], 11 million malware variants were discovered by 2008 and 90% of this malware comes from hidden downloads from popular and often trusted websites. Before a download takes place, a user is first required to visit the malicious site. To lure the user into visiting a website with malicious content, attackers would send spam emails that contain links to the site. When unsuspecting user visits the malicious website, malware is downloaded and installed in the victim's machine without the knowledge of the user. For example, the infamous Storm worm makes use of its own network, multiple of infected computers, to send spam emails containing links to such attack pages.

Emerging Threats in cyber security

Cyber-attacks on cyberspace evolve through time capitalizing on new approaches. Most times, cyber criminals would modify the existing malware signatures to exploit the flaws exist in the new technologies. In other cases, they simply explore unique characteristics of the new technologies to find loopholes to inject malware. Taking advantages of new Internet technologies with millions and billions active users, cyber criminals utilize these new technologies to reach out to a vast number of victims quickly and efficiently. We select four such up and coming technology advancements which include: social media, cloud computing, smartphone technology, and critical infrastructure, as illustrative examples to explore the threats in these technologies. We discuss unique characteristics of each of these emerging technologies and analyse a number of common attack pattern.

A. Social media

Social media, such as Facebook and Twitter, has shown explosive growth in recent years. At the end of 2012, there are more than 450 million active user accounts in Twitter while the number grows exponentially in Facebook reaching almost.1 billion users. Social networking sites have been very popular and become the preferred method of communication for most young generations. Each of these social media websites typically provide tools where users share their personal information (i.e., name, address, gender, date of birth, preference in music and movie), photos, stories and disseminate links. Attackers are taking advantage of the social media craze as a new medium for launching insidious attacks. By the end of 2008, the Kaspersky Lab collection contained more than 43,000 malicious files relating to social media sites. A report published by IT security and data protection firm Sophos has revealed an alarming rise in attacks on users of social media websites. According to their report, around 60% of the users in the social networks have received spam. Due to the unlimited access to the profile of users, attackers can further gain the information of corporation and commercial secrets. In the survey conducted by Sophos , around 60% companies concern that their employees provide too much information in social networks while around 66% companies think that using social networks pose a great threat to the companies.

Koobface worm [8] that spreads through social media sites in 2009 is notably the best-known malware case that utilizes the proliferation of social media sites. Leveraging its zombie arsenal, the Koobface botnet automates the creation of new social media accounts used to befriend unsuspecting users, in turn spamming enticing links that redirect to malware. Victims that fall prey to the social engineering attacks witness their own social networking accounts turn into vehicles for sending spam to the victim's friends, while the victim's machine is repurposed into a zombie. Thomas and Nicol [8] constructed a zombie emulator which was able to infiltrate the Koobface botnet and identified fraudulent and compromised social network accounts used to distribute malicious links to over 213,000 social network users generating over 157,000 clicks. They discovered the ineffectiveness of current blacklisting services offered by social network operators to filter malicious virus through most prominent blacklisting services. They argued that those blacklisting services only recognize 27% of threats and take on average 4 days to respond while they found that 81% of visitors to Koobface's spam occur within the first 2 days of a link being posted, leaving the majority of social networking users vulnerable. Another popular malware attack is done by the use a significant number of Twitter or Facebook accounts that are not legitimate or not in use. Cyber criminals are becoming a lot more sophisticated in their efforts to appear as trustworthy users. Then the criminals trick users in the social network site into "friendling" or following them and clicking on their status updates which often lead to malicious web sites. In another study [109], it is illustrated that a large number of malwares were spread after clicking for content on "trending" topics via Twitter. Understanding the social network platforms and simulations to spread malware using mock up services over Facebook has also been studied in [9].

Social networking sites also have raised the stakes for privacy protection because of the centralization of massive amounts of user data, the intimacy of personal information collected, and the availability of up-to-date data which is consistently tagged and formatted [9]. This makes social networking sites an attractive target for a variety of organizations seeking to aggregate large amounts of user data, some for legitimate purposes and some for malicious ones. In most cases, extracting data violates users' expectation of privacy. Protecting user's private data kept in the social networking service providers has been explored. Lucas et al [10] proposed a Facebook application for encrypting and decrypting sensitive data using client-side JavaScript. This architecture ensures that data never arrives at the social network service providers in an unencrypted form preventing them from observing and accumulating the information that users transmit through the network. Privacy awareness related issues and tools which can help users to set their privacy setting more intuitively have been proposed as well. For example, Fang and LeFevre [1] proposed privacy wizard. The wizard iteratively asks the user to assign privacy "labels" to selected friends, and it uses this input to construct a classifier, using a machine learning model, which can in turn be used to automatically assign privileges to the rest of the user's friends. The intuition for the design comes from the observation that real users conceive their privacy preferences of which friends should be able to see which information, based on implicit set of rules they set and repeatedly use in most friends setting.

B. Cloud computing

The efficiencies of moving data and applications to the cloud continue to attract consumers who store their data in DropBox and iCloud, use Gmail and Live mail to handle email, and track their lives using services such as Evernote and Mint.com. Cloud computing is arguably one of the most significant technological shifts in recent times [16]. The mere idea of being able to use computing in a similar manner to using a utility is revolutionizing the IT services world and holds great potential. Customers, whether large enterprises or small businesses, are drawn towards the cloud's promises of agility, reduced capital costs, and enhanced IT resources. IT companies are shifting from providing their own IT infrastructure to utilizing the computation services provided by the cloud for their information technology needs.

Cloud computing provides unique characteristics that are different from the traditional approaches. The five key characteristics of cloud computing include on-demand self-service, ubiquitous network access, location-independent resource pooling, rapid elasticity, and measured service, all of which are geared towards using clouds seamlessly and transparently [67]. Resource polling refers to the ability where no resources are dedicated to one user but instead are pooled together to serve multiple consumers. Resources, whether at the application, host or network levels, are assigned and reassigned as needed to these consumers. On demand self-service refers where the users can assign themselves additional resources such as storage or processing power automatically without human intervention. This is comparable with autonomic computing where the computer system is capable of self-management. Along with self-provisioning of resources, cloud computing is characterized with the ability to locate and release resources as rapidly as needed, the term often called as "elasticity". This allows consumers to scale up the resources they need at any time to address heavy loads and usage spikes, and then scale down by returning the resources to the pool when finished. Measured service, also often called as pay as you go, enables the cloud to be offered as a utility where users pay on a consumption basis, much the same way it is done to pay utilities like electricity, gas and water.

Cloud computing is also a model of integration that delivers various resources to clients at different layers of the system and utilizes different resources. Generally speaking, the architecture of a cloud computing environment can be divided into 4 layers: the hardware layer (including data centres), the infrastructure layer, the platform layer and the application layer.

- The hardware layer: This layer is responsible for managing the physical resources of the cloud, including physical servers, routers, switches, power and cooling systems. In practice, the hardware layer is typically implemented in data centres. A data centre usually contains thousands of servers that are organized in racks and interconnected through switches, routers or other fabrics. Typical issues at hardware layer include hardware configuration, fault tolerance, traffic management, power and cooling resource management.
- The infrastructure layer: This layer is also known as the virtualization layer. The infrastructure layer creates a pool of storage and computing resources by partitioning the physical resources using virtualization technologies such as Xen,

Kernel based Virtual Machine and VMware. The infrastructure layer is an essential component of cloud computing, since many key features, such as dynamic resource assignment, are only made available through virtualization technologies.

- The platform layer: Built on top of the infrastructure layer, the platform layer consists of operating systems and application frameworks. The purpose of the platform layer is to minimize the burden of deploying applications directly into VM containers. For example, Google App Engine operates at the platform layer to provide API support for implementing storage, database and business logic of typical web applications.
- The application layer: At the highest level of the hierarchy, the application layer consists of the actual cloud applications. Different from traditional applications, cloud applications can leverage the automatic-scaling feature to achieve better performance, availability and lower operating cost.

However, in practice, clouds offer services that can be grouped into three categories: software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS) [68]. Applications running on or being developed for cloud computing platforms pose various security and privacy challenges depending on the underlying delivery and deployment models. In IaaS, the cloud provider supplies a set of virtualized infrastructural components such as virtual machines (VMs) and storage on which customers can build and run applications. The application will eventually reside on the VM and the virtual operating system. PaaS enables programming environments to access and utilize additional application building blocks. Such programming environments have a visible impact on the application architecture, such as constraints on which the application can request services from an OS. Finally, in SaaS, the cloud providers enable and provide application software as on-demand services.

Multi-tenancy is a feature unique to clouds which allows cloud providers to manage resource utilization more efficiently by partitioning a virtualized, shared infrastructure among various customers. For example, to isolate multiple tenants' data, Salesforce.com employs a query rewriter at the database level, whereas Amazon uses hypervisors at the hardware level. Virtualization is an important enabling technology in this area that helps abstract infrastructure and resources to be made available to clients as isolated VMs. Providing strong isolation, mediated sharing, and secure communications between VMs are active research areas. Using a flexible access control mechanism that governs the control and sharing capabilities of VMs within a cloud host has been suggested as a potential solution. Because clients acquire and use software components from different providers, crucial issues include securely composing them and ensuring that information handled by these composed services is well protected. For example, a PaaS environment might limit access to well-defined parts of the file system, thus requiring a fine-grained authorization service.

Trust management and policy integration is an active area of research in cloud computing as the outsourcing model of the cloud, where the cloud providers control and manage user's data and services, forces the clients to have significant trust in their provider's technical competence. In cloud computing environments, the interactions between different service domains driven by service requirements are also dynamic, transient, and intensive. Thus, a development of trust framework has been proposed to allow efficient capturing of a generic set of parameters required for establishing trust and to manage evolving trust and interaction/sharing requirements. The cloud's policy integration is another active area of research to address challenges such as semantic heterogeneity, secure interoperability, and policy-evolution management. Furthermore, customers' behaviours can evolve rapidly, thereby affecting established trust values. This suggests a need for an integrated, trust-based, secure interoperation framework that helps to establish, negotiate, and maintain trust to adaptively support policy integration.

C. Smartphones

Smartphones, coupled with improvement in wireless technologies, have become an increasingly sophisticated computer and communication device that is readily carried by individuals throughout the day. The convergence of increasing computing power, personalization and mobility makes them an attractive means of planning and organizing work and private life of individuals. According to [16], the sheer volume of mobile phone users around the world indicates a current need for proactive mobile security measures. It is assumed that over 4.5 billion use a cell phone every day and an estimated 2 billion smartphones will be deployed by 2013.

There are a number of different styles of attacks targeted to take advantage of the proliferation of mobile computing. Communication related attacks are derived from flaws in the design and management of mobile communication infrastructure. The attacker may try to break the encryption of the mobile network. The GSM (Global System for Mobile Communication) network today uses two variants of algorithms known as A5/1 and A5/2, latter being known to be weaker. Since the encryption algorithm was made public, it was proved that it is possible to break the encryption in about 6 hours [118]. An attacker can try to eavesdrop on Wi-Fi communications to derive information (e.g. username, password). These types of attacks are not unique to smartphones, but they are very vulnerable to these attacks because very often the Wi-Fi is the only means of communication, they have to access the Internet. Security issues related to Bluetooth on mobile devices have been studied and have shown a number of problems. The worm searches for nearby phones with Bluetooth in discoverable mode and sends itself to the target device. The user must accept the incoming file and install the program. After installing, the worm infects the machine. To prevent communication related attacks, network traffic exchanged by phones can be monitored such as surveillance on network routing points or monitoring the use of network mobile protocols.

Another type of attacks is derived from the vulnerabilities in mobile software applications especially exploiting mobile web browser. Just as common Web browsers, mobile web browsers are extended from pure web navigation with widgets and plug-ins which many attackers use as means to spread malware through. Jailbreaking the iPhone was based entirely on vulnerabilities on the web browser based on a stack-based buffer overflow in a library used by the web browser.

Malicious attackers target mobiles phones as a medium to spread malware [8]. Both Georgia Tech emerging cybersecurity threats reports and Symantec threats reports in last couple of years warn the growing number of malwares that are specifically created for mobile phones such as targeting Google Android based phones and Apple iPhones. To control the malware propagation, mobile companies offer a centralized public market place complimented with an approval process before hosting the application. The centralized marketplace helps to remove any application if found suspicious before they are downloaded by the users. For example, Apple adopts a vetting process to ensure all applications conform to Apple's rules before they can be offered via the App Store. Apple approves an application by code signing with encryption keys. Accessing the applications via

App store is the only way for iPhone devices to install applications. Similar to Apple, Android too has a public marketplace to host applications. However, unlike Apple, the Android application can be self-signed. Android uses crowd sourcing to rate the applications by users. Based on user complaints, applications can be removed from marketplace and remove them from the device as well. Another approach taken by the mobile companies to protect their mobile platforms found in the idea of a sand boxing. Sand boxing compartmentalizes different processes to prevent them from interacting and damaging each other therefore effectively limiting any chance for malicious code to be implanted and overtaking the running processes from doing harmful activities. Apple iOS focuses on limiting access to its API for applications from the Apple Store while Android uses its sandboxing on underlying legacy Linux kernel.

C. Critical infrastructure

The critical infrastructure systems that form the lifeline of a modern society and their reliable and secure operation are of paramount importance to national security and economic vitality. In most sense, the cyber system forms the backbone of a nation's critical infrastructures, which means that a major security incident on cyber systems could have significant impacts on the reliable and safe operations of the physical systems that rely on it. The recent findings, as documented in government reports, indicate the growing threat of physical and cyber-based attacks in numbers and sophistication on electric grids and other critical infrastructure systems.

Critical infrastructure protection is harder to address than information and communication technology (ICT) protection because of these infrastructures' interconnection complexity, which can lead to different kinds of problems. Consider the power grid, in which geographically dispersed production sites distribute power through different voltage level stations (from higher to lower voltage) until energy eventually flows into our houses. Both the production and distribution sites are typically controlled by supervisory control and data-acquisition (SCADA) systems, which are remotely connected to supervision centres and to the corporate networks (intranets) of the company's managing the infrastructures. The intranets are linked to the Internet to facilitate, for example, communication with power regulators and end clients. These links create a path for external attackers. Operators' access SCADA systems remotely for maintenance operations, and sometimes equipment suppliers keep links to the systems through modems. The prevalence of proprietary solutions and use of older versions plagued with vulnerabilities are sought to add another dimension to propose solutions to protect nation's crucial infrastructure.

D. Other emerging areas of concern

Cybersecurity in embedded systems and sensors are the topics that have received an increasing amount of attention from industry and academia in recent years due to their increased use in every facet in our lives. For example, embedded small devices inserted in cars, home appliances, mobile phone, and audio/video equipment's, increasingly become a part of our lives. Similarly, sensors are seeing broader research and commercial deployments in military, scientific, and commercial applications including monitoring of biological habitats, agriculture, and industrial processes. Security concerns in these areas are different from the traditional security problems in PC and enterprise computing due to their different embedded nature and operational environment. Embedded systems and sensors are often highly cost sensitive requiring them to use smaller processors which have limited room for security overhead for example storing a big cryptography key. Therefore, the most enterprise security solutions do not work in the embedded system world. Embedded systems and sensors are resource constrained in energy, memory, computational speed and communications bandwidth due to the nature of small size. They have a very weak physical trust boundary. For example, they are installed in residents and commercial properties, outside fields, or carried by human in their hands or pockets which enables many different physical-oriented attacks. They use an intimate connection between hardware and software often without the shielding of an operating system. The different embedded nature of the embedded systems and sensors have created different sets of security vulnerabilities. For example, limited battery power in embedded systems makes them vulnerable to attacks that drain this resource. The proximity of embedded systems to a potential attacker creates vulnerabilities for attacks where physical access to the system is necessary. This allows the attackers to perform attacks that are involved examining the usage of physical system, for example, power analysis attacks or snooping attacks on the system bus. Embedded systems need to operate within a reasonable environmental condition. Due to the highly exposed operating environment of embedded systems, there is a potential vulnerability for attacks that overheat the system (or cause other environmental damage). Attackers reprogram a stolen embedded system to use them for further misuse. The usual security countermeasures to prevent unauthorized access through user authentication, techniques to preserve data integrity through cryptographies and network defence mechanisms are active area of interest in the field. However, preventing attacks done by examining or altering the physical system are quite unique, for example techniques such as masking, window methods and dummy instruction insertion in the code/algorithm have been proposed. Since Network connectivity via wireless or wired access is increasingly common for embedded systems to enhance remote control data collection and update, the vulnerabilities that exploiting such network connectivity, such as spread of viruses and wiretapping, have become another source of growing concern in the field.

Cyber warfare refers to politically motivated hacking to conduct sabotage and espionage. In the book Cyber Ware, cyber warfare was defined as "actions by a nation-state to penetrate another nation's computers or networks for the purposes of causing damage or disruption". Most cyber warfare concerns are focused on national security breaches and sabotage of nation's critical infrastructure. The former case concerns with the international espionage where classified information that may breach the national security is illegally accessed or altered by unauthorized people. The latter case concerns with any potential disruption of nation's critical infrastructure such as power grid system and transportation system. In 2008, a simulated exercise codenamed "cyber storm" was conducted by the Department of Homeland Security. The purpose of the exercise was to test the nation's defense against digital espionage. The Cyber Storm exercise highlighted the gaps and shortcomings of the nation's cyber defenses. Since then, researchers have proposed a number of new priorities in regard to nation's cyber defense strategy. Identification of Nation's Critical Systems has been proposed [80] to recognize the Internet-enabled systems that are critical to nation's cyber defense and any interdependencies among the systems. A number of strategies to protect nation's critical infrastructures by vulnerability identification and remediation, and threats mitigation and response have been proposed.

III. ML AND DMMETHODS FOR CYBER SECURITY

This section describes the different ML/DM methods for cyber security. Each technique is described with some detail, and references to seminal works are provided. Also, for each method, two to three papers with their applications to cyber domain are presented.

A. Artificial Neural Networks

ANNs are inspired by the brain and composed of inter-connected artificial neurons capable of certain computations on their inputs [22]. The input data activate the neurons in the first layer of the network whose output is the input to the second layer of neurons in the network. Similarly, each layer passes its output to the next layer and the last layer outputs the result. Layers in between the input and output layers are referred to as hidden layers. When an ANN is used as a classifier, the output layer generates the final classification category.

B. Association Rules and Fuzzy Association Rules

The goal of Association Rule Mining is to discover previously unknown association rules from the data. An association rule describes a relationship among different attributes: IF (A AND B) THEN C. This rule describes the relationship that when A and B are present, C is present as well. Association rules have metrics that tell how often a given relationship occurs in the data. The support is the prior probability (of A, B, and C), and the confidence is the conditional probability of C given A and B. Association Rule Mining was introduced by Agrawal et al. [29] as a way to discover interesting co-occurrences in supermarket data. It finds frequent sets of items (i.e., combinations of items that are purchased together in at least N transactions in the database), and from the frequent items sets such as {X, Y}, generates association rules of the form: $X \rightarrow Y$ and/or $Y \rightarrow X$. A simple example of an association rule pertaining to the items that people buy together is:

IF (Bread AND Butter) → Milk (1)

This rule states that if a person buys bread and butter, they also buy milk.

C. Bayesian Network

A Bayesian network is a probabilistic graphical model that represents the variables and the relationships between them. The network is constructed with nodes as the discrete or continuous random variables and directed edges as the relationships between them, establishing a directed acyclic graph. The child nodes are dependent on their parents. Each node maintains the states of the random variable and the conditional probability form. Bayesian networks are built using expert knowledge or using efficient algorithms that perform inference.

D. Clustering

Clustering is a set of techniques for finding patterns in high-dimensional un labelled data. It is an unsupervised pattern discovery approach where the data are grouped together based on a similarity measure. The main advantage of clustering for intrusion detection is that it can learn from audit data without requiring the system administrator to provide explicit descriptions of various attack classes.

E. Decision Trees

A decision tree is a tree-like structure that has leaves, which represent classifications and branches, which in turn represent the conjunctions of features that lead to those classifications. An exemplar is labelled (classified) by testing its feature (attribute) values against the nodes of the decision tree. The best-known methods for automatically building decision trees are the ID3 [12] and C4.5 algorithms. Both algorithms build decision trees from a set of training data using the concept of information entropy. When building the decision tree, at each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of examples into subsets. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then performs recursion on the smaller subsets until all the training examples have been classified. instance with a trained model, and understandability of the final solution (classification) of each ML or DM method.

Depending on the particular IDS, some might be more important than others.

IV.CONCLUSION

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.

V. REFERENCES

[1]. A. Mukkamala, and A. Sung, and A. Abraham, “Cyber security challenges: designing efficient intrusion detection systems and antivirus tools,” Vemuri, V. Rao, Enhancing Computer Security with Smart Technology. (Auerbach, 2006) (2005), pp. 125–163

[2]. M. Bhuyan, D. Bhattacharyya, and J. Kalita, “Network anomaly detection: Methods, systems and tools,” IEEE Communications Surveys & Tutorials, vol. PP, no. 99, 2013, pp. 1–34

[3]. E.E. Schultz Where have the worms and viruses gone? New trends in malware Comput. Fraud Secur., 2006 (7) (2006), pp. 4-8

[4]. U. Bayer, I. Habibi, D. Balzarotti, E. Kirda, C. Kruegel, A view on current malware behaviours, in: USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET), April 2009.

[5]. G. Cluley Sizing up the malware threat-key malware trends for 2010 Netw.Secur. (2010), 10.1016/S1353-4858(10)70045-3

[6]. L.F. Cranor, S. Garfinkel Secure or usable? IEEE Secur. Priv., 2 (2004), pp. 16-18

[7]. Anti phishing group tech reports: http://www.antiphishing.org/phishReportsArchive.html,lastaccessed:June2013.

[8]. K. Thomas, D. M. Nicol, The Koobfacebotnet and the rise of social malware, in : Proceeding softhe 5th International Conferenceon Maliciousand Unwanted Software (Malware2010), 2010, pp. 63–70.

[9]. W.Luo, J. Liu, J. Liu, C. Fan, Ananalysis of security in social networks, in: Eighth IEEE International Conferenceon Dependable, Autonomic and Secure Computing, 2009, pp. 648–651

[10]. M. Lucas, N. Borisov, Flybynight : Mitigating the privacy risks of socialnetworking, in: WPES,2008.

[11]. L. Fang, K. LeFevre, Privacywizards for social networking sites, in : Proc. WWW’10,pp.351–360.

[12]. A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," in IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, Secondquarter 2016, doi: 10.1109/COMST.2015.2494502.

[13]. A. Siffer, P. -A. Fouque, A. Termier and C. Largouet, "Netspot: A Simple Intrusion Detection System with Statistical Learning," 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, 2020, pp. 911-918, doi: 10.1109/TrustCom50675.2020.00122.

Cite this article as :

Ashish Prajapati, Shital Gupta, "A Survey : Data Mining and Machine Learning Methods for Cyber Security ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 2, pp. 24-34, March-April 2021. Journal URL : https://ijsrcseit.com/CSEIT217212