Study of Security Behavior Using Data Mining Techniques: A Review

Mohammad Iquebal Akhter [1]
Research Scholar [1]
Prof. Mohammed Gulam Ahamad [2]
Department of Computer Engineering, Sphoorthy Engineering College, Hyderabad [1, 2]

Abstract- *Data mining technology has been used mainly for the past ten years to use the security behavior of data mining. As innovation develops, the battle between security analysts and malware scholars is timeless. The proposed approach is not enough, and the evolution and complex nature of malware is rapidly changing and therefore more difficult to identify. This paper systematically and in detail studies on the use of data mining techniques for security behaviors of data through machine learning. Safe behavior methods have been compared to one another based on their importance. Their advantages and disadvantages are discussed from the perspective of data mining models, assessment methods and proficiency. The survey helps researchers have a comprehensive understanding of the areas of malware testing and subsequent inspections by experts.*

Keywords - Data Mining, Malware, Security, Secure Behavior Technology, Information Discovery Task.

I. Introduction

Data mining is also the process of automatically discovering the necessary information in a large database. Data mining technology is positioned to find useful and novel patterns that are still unknown. They also provide the need to predict future observations. Not all information discovery tasks are considered data mining. For example, a task similar to an information retrieval area is to find a particular web page by querying an Internet search engine or using a database management system to find a single record. While these tasks are necessary and may involve data structures and complex algorithms, they do use them on traditional computer science techniques and obvious data features to develop index structures for efficiently retrieving and organizing information. Despite this, data mining techniques have been used to enhance information retrieval systems.

1.1 Data mining methods

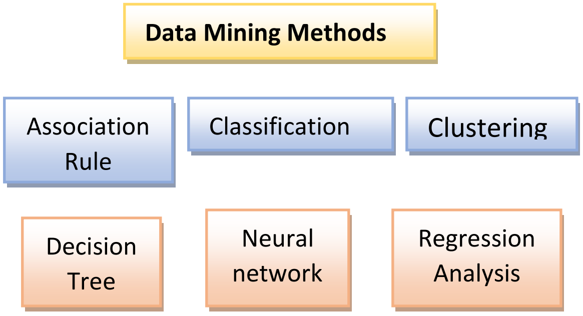

Fig 1.1 Data Mining Methods

Many methods are used to mine big data, but the following eight are the most common. Association rules help to find possible relationships between variables in the database, discover hidden patterns, and identify variables and how often they occurrence.

II. Classification breaks a large dataset into predefined classes or groups.

Clustering helps identify data items with similar characteristics and understands similarities and differences between data. Decision tree technology creates classification and regression models in the form of tree structures. Neural network technology is used to simulate the complex relationship between input and output. Discover new patterns. Regression analysis is used to predict the value of an item based on a known value of other items in the data set by constructing a model that relates the relationship between the variable and the independent variable. Statistical techniques help to find patterns and build predictive models. Visualize the new patterns and display the results in a user understandable way.

This paper has been propose on the large amounts of data sets, from simple digital measurements and text documents to more complex information such as spatial data, multimedia channels and hypertext documents. The following is a non-exclusive list of various information collected in digital form in database and flat files.

- Business transactions: Every transaction in the business industry (usually) is “remembered” as permanent. Such transactions are typically time dependent and may be inter-business transactions such as purchases, exchanges, banks, stocks, etc., or intra-business transactions such as management of internal goods and assets. For example, due to the widespread use of barcodes, large department store millions of transactions per day, representing often terabytes of data. Storage space is not a major issue, as the price of hard drives is declining, but the effective use of data for competitive decision making within a reasonable time frame is undoubtedly the most important issue for companies to survive. A highly competitive world.

• Scientific Data: whether calculating particles in the Swiss nuclear accelerator laboratory, studying the readings of grizzly radios in Canadian forests, collecting data on marine activities on Antarctic icebergs, or investigating human psychology at American universities, society is accumulating a large amount of scientific data that needs to be analyzed. Unfortunately, we can capture and store more new data faster than analyzing old data that has accumulated.

• Medical and personal data: From government censuses to people and client files, we continuously collect a wealth of information about individuals and groups. Organizations such as governments, companies and hospitals are storing very important personal data to help them manage their human resources, better understand the market or simply help customers. Regardless of the privacy issues that such data often shows, this information is collected, used, and even shared. This information can reveal customer behavior and the like when related to other data.

• Surveillance videos and pictures: With the amazing collapse of camera prices, cameras are becoming ubiquitous. The tape from the surveillance camera is usually recycled, so the content is lost. However, there is a trend today to store tapes and even digitize them for future use and analysis.

• Satellite sensing: There are countless satellites in the world: some are located above the area, some are in Earth orbit, but are sending uninterrupted data streams to the ground. NASA, which controls a large number of satellites, receives more data per second than any NASA researchers and engineers can process. Many satellite images and data will be released as soon as they are received, and will be analyzed by other researchers.

• Games: Their society is collecting a lot of data and statistics about games, athletes and athletes. From hockey scores, basketball passes and racing mistakes, to swimming time, boxer advancement and chess position, and all data is stored. Commentators and journalists are using this information for reporting, but trainers and athletes want to use this data to improve performance and better understand their opponents.

• Digital media: The popularity of inexpensive scanners, desktop cameras and digital cameras is one of the reasons for the explosive growth of digital media storage. In addition, many radio stations, TV channels and movie studios are digitizing their audio and video collections to improve the management of their multimedia assets. Associations such as the NHL and the NBA have begun to convert their vast collection of games into digital form.

• CAD and software engineering data: There are many computer-aided design (CAD) systems for architects to design buildings or engineers to conceptualize system components or circuits. These systems are generating large amounts of data. In addition, software engineering is a source of similar data, such as code, libraries, objects, etc., and requires powerful management and maintenance tools.

•Virtual World: There are many applications that take advantage of 3D virtual spaces. These spaces and the objects they contain are described in a special language such as VRML. Ideally, these virtual spaces are described in such a way that they can share objects and locations. There are a large number of virtual reality objects and spatial repositories available. Managing these repositories and performing content-based searches and retrievals from these repositories remains a research issue, and the scale of collection continues to grow.

• Text reports and memos (e-mails messages): Most communications with in and between companies or research organizations and even private individuals are based on reports and memos in the form of text that is usually exchanged by e-mail. These messages are regularly stored in digital form for future use and reference to create a powerful digital library.

•The World Wide Web Repositories: Since the birth of the World Wide Web in 1993, documents of various formats, content, and descriptions have been collected and interconnected with hyperlinks, making it the largest data repository ever. Despite the dynamic and unstructured nature of the World Wide Web, its heterogeneous nature, and often redundancy and inconsistency, the World Wide Web is the most commonly used reference dataset due to the wide variety of topics covered, resources and publisher contributions. Many people think that the World Wide Web will become a compilation of human knowledge.

III. Background

[1] Kantarcıoglu & Vaidya study the horizontal partitioning data of the Naive Bayes classifier for privacy protection. According to them, the issue of secure distributed classification is an important issue. In many cases, the data is divided into several organizations. These organizations may wish to use all data to create more accurate predictive models without revealing their training data/foundation or classification. The Naive Bayes classifier is a simple but effective classifier. In this paper, they present a kind of confidentiality that keeps the Naive Bayes classifier horizontally partitioned.

The results of this paper also support the view that there are several useful security protocols that can safely implement many distributed data extraction algorithms. Currently, it has developed such a set of tools. As part of their future work, they plan to conduct experiments to verify the feasibility and scalability of the method. Although there are general techniques to extend any of the security algorithms in the semi-honest model to resist a certain degree of collusion, an effective solution to their goals can be designed. They plan to continue to explore this direction in the future. Given the growing focus on privacy, more data processing algorithms need to be adjusted to maintain confidentiality.

[2] Panackal & Pillai studies on privacy protection data mining. According to them, mining data applications involve data-rich environments such as biomedical, electronic health records, the Internet, web logs and wireless

networks, mobile data from sensors and many others. Confidentiality in the data extraction process can be achieved by various techniques such as interference and encryption. This paper attempts to reaffirm the various data protection technologies (PPDM) currently being developed to address the confidentiality issues in data mining. Studies have shown that when developing a common solution, the compromise between confidentiality and loss of information creates a bottleneck. This paper explores various PPDM techniques based on the hierarchical structure of PPDM classification, and finally ends with several research directions in the future.

The main focus of this paper is to reiterate the various PPDM techniques in the literature to manage privacy issues in data utilization. Most studies have shown trade-offs between privacy, information loss, and computing expenses. Maximizing the usefulness of data by storing information is a key challenge while protecting confidentiality. In order to provide accurate results in the data extraction process, many PPDM technologies are task-based. Since no such technology transcends all confidentiality issues, research in this area can make a significant contribution. This research can be done using any prior art or a combination of them, as shown or by fully developing new technologies. Some interesting PPDM frames are shown and can still be used. This survey will certainly help researchers set their own privacy goals based on specific requirements.

[3] Wahlstrom and Roddick (2001) argue that knowledge discovery and data mining are powerful tools for automated data analysis and are expected to become the most commonly used analysis tools in the near future. The rapid spread of these technologies requires an urgent review of their social impact. This paper identifies social problems caused by knowledge discovery (KD) and data mining (DM). An overview of these techniques is presented, followed by a detailed discussion of each issue. The purpose of this paper is to first explain the cultural background of each issue, and secondly, to describe the impact of KD and DM in each case. Existing solutions to each problem are identified and checked for feasibility and effectiveness, and a solution that provides reasonable context-sensitive means for collecting and analyzing data is presented and briefly presented. The discussion of a topic is discussed at the end of this article.

[4] Thuraisingham (2002) studies data mining, national security, privacy and civil liberties. In this article, they describe privacy threats that may occur through data mining, and then treat privacy issues as a variant of the reasoning problem in the database. This work is dedicated to important privacy areas related to networking and data utilization. Although data mines have been tried to address national security and information security issues such as intrusion detection, they have focused on the negative impact of data mines. In particular, they discussed inferred issues that may arise from exploitation, as well as ways to violate privacy, especially due to access to network data. First, they outlined the reasoning problem and then discussed ways to solve this mining problem. Warehousing and reasoning issues are also discussed. Then they outlined the privacy issue. Then they discussed how the inference controller handles privacy and analyzed the mining of sensitive data. Finally, they discussed civil liberties regarding national security. Although there are few reports on the Web and mining privacy issues, they are moving in the right direction. People are becoming more aware of the problem, and organizations such as the IFIP Database Security Working Group have made it a priority. Some progress is expected in terms of research projects in this area. Keep in mind that there are still some social and political issues to consider. Technical experts, sociologists, policy experts, counter-terrorism experts and legal experts must work together to fight terrorism and ensure privacy. Keep in mind that the number of cyber security conferences will increase, including confidentiality workshops. Recently, FSN sponsored a next-generation data mining seminar on counter-terrorism and privacy. As the Internet becomes more complex, there are more and more threats. Therefore, they must always be vigilant and continue to investigate, design and implement different online privacy measures, but at the same time they need to ensure national security.

[5] Kantarcıoglu et al. (2004) Privacy-protected data mining focuses on obtaining valid results when data entry is private. An extreme example is the method based on Secure Multiparty Computation, in which only the results are displayed. This article examines this problem by developing a framework that addresses this issue. Present metrics and analyses where these values are consistent with obvious issues.

Increasing the power of computing resources and ubiquity is a constant threat to personal privacy. Data tools that maintain data confidentiality and multi-party secure computing make data processing possible without disclosure, but without privacy issues. They have defined this problem and explored how data mining can be used to undermine privacy. They gave definitions to simulate the impact of data mining on privacy, analyzed they definition of Gaussian mixture for two types of problems, and presented a heuristic example that could be applied to more general scenarios.

[6] Brickell & Shmatikov studied the destruction of data mining utilities in anonymized data distribution. In this article, they ask whether the generalization and suppression of quasi-identifiers provide any benefit over simple cleaning. Simply distinguishes quasi-identifiers from sensitive ones. Previous work has shown that k-anonymous databases can be used for data mining, but k-anonymization does not guarantee any privacy. In contrast, they measure the trade-off between privacy (how much can the opponent learn from cleaned up records) and utility, and measure the accuracy of the data mining algorithm on the same cleanup record. For their experimental evaluation, the same data set from the UCI machine learning library was the same as used in previous studies on generalization and inhibition. Their results show that even moderate privacy gains require almost complete destruction of data mining utilities. In most cases, trivial cleanup provides the same utility and better privacy as k-anonymity, - diversity and similar methods based on generalization and suppression.

Micro data privacy can be understood as preventing membership from being exposed (the adversary should not know if a particular individual is included in the database) or sensitive properties being exposed (the cleaned database should not display a lot of information about any personally sensitive attributes). As we all know, generalization and suppression cannot prevent the disclosure of membership. For sensitive property disclosure, perfect privacy can be achieved by simply removing sensitive attributes or quasi-identifiers from published data to achieve a very weak opponent, only knowing the quasi-identifier. Of course, these trivial clean ups will also break any utilities that rely on the delete attribute.

[7] Gkoulalas & Verykios (2009) studied an overview of privacy protection data mining. In this article, they briefly introduce confidential data protection, which is one of the most popular guides in the data mining research community. In the first part of this article, they highlighted the need for this area of research and its many applications. They discussed different categories of methods proposed to protect data-sensitive data itself or the powerful data results of data mine applications. Finally, they provided some road maps based on the most promising future directions in the region.

[8] Friedman & Schuster studied on Data Mining with Differential Privacy. They consider the mining issue with official privacy safeguards, considering a data access interface based on the confidential privacy framework. Differential confidentiality implies that calculations are not susceptible to changes in records of a particular individual, thus limiting data leakage through results. The Privacy Interface ensures uninterrupted access to data and does not require any confidentiality expertise from the data mining operator. However, as they have shown in the paper, naive use of the interface to build confidentiality data storage algorithms that could preserve data confidentiality could lead to lower data mining results. They solve this problem while considering privacy and algorithm requirements, focusing on inducing the decision tree as a test application. The Privacy Mechanism has a profound effect on the performance of the data mining methods. They demonstrate that this choice could make a difference between a precise classifier and a totally useless one. Moreover, an improved algorithm can achieve the same level of precision and confidentiality as naive implementation, but with a lesser order of learning.

[9] Chris Clifton (2001) studied distributed data mining with privacy protection. This obtaining effective data mines in distributed data sets and a limited desire to share data between sites. They recommend performing local operations on each site to generate intermediate data that can be used to obtain results without revealing private information for each site. There are many variations of this problem, depending on how the data is distributed, the type of data mining they want to do, and the restrictions on information exchange. Some problems are easy to handle and others are more difficult to deal with.

[10] Clifton studied the security and privacy implications of data mining. As we all know, data extraction allows us to find information that can only be found in the database. This could be a security/privacy issue. This position paper discusses possible problems and solutions and presents ideas for future research in this area.

IV. Detection process

When using data mining, malware detection involves two steps:
Extraction function
Classification/cluster

Extraction features
Data Mining Malware Detection
Classifying / Clustering

Fig 2: Data Mining Malware Detection

In the first step, various functions such as API calls, n-grams, binary strings, and program behavior are statically and dynamically extracted to capture the characteristics of the file samples. Feature extraction can be performed by running static or dynamic analysis (without actually running potentially harmful software). A hybrid approach combining static and dynamic analysis can also be used. During classification and clustering, file samples are classified into groups based on feature analysis. To categorize your samples, it can use classification or clustering techniques.

To classify file samples, this need to use a classification algorithm (such as RIPPER, Decision Tree (DT), Artificial Neural Network (ANN), Naive Bayes (NB) or Support Vector Machine (SVM) to build the classification model (classifier) Clustering is used to group malware samples with similar characteristics. Using machine learning techniques, each classification algorithm builds a model that represents both benign and malicious classes. Training the classifier with such a collection of file samples makes it possible to detect newly released malware. It can apply one or more data mining methods to create an effective model to ensure successful detection of the attack.

V. Conclusion

This paper focuses on data mining techniques and internal intrusion detection and protection forensics techniques. The proposed system enables data mining and security defense system based on machine learning to identify system calls, create user profiles and isolate them from attacker profiles to protect users from internal attacks.

References

[1] Kantarcıoglu, M. & Vaidya, J. Privacy Preserving Naive Bayes Classifier for Horizontally Partitioned Data.
[2] Panackal, J.J, & Pillai, A.S. (2013). Privacy Preserving Data Mining: An Extensive Survey. *Proc. of Int. Conf. on Multimedia Processing, Communication and Info. Tech., MPCIT, 297-298.*
[3] Wahlstrom, K. & Roddick, J.F. (2001).On the Impact of knowledge Discovery and Data Mining.22-27.
[4] Thuraisingham, B. Data Mining, National Security, Privacy and Civil Liberties. Volume 4, (Issue 2), 1-5.
[5] Kantarcıoglu, M., & Jiashun, J., Clifton, J.C. (2004). When do Data Mining Results Violate Privacy. *599-604.*

[6] Brickell, J.J., &Shmatikov, J.V. (2008). The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing. 70-78.

[7] Divanis, G.G, &Verykios. V.S. (2009). An Overview of Privacy Preserving Data Mining. Vol. 15, No. 4. 23-26.

[8] Friedman, A. A, & Schuster. A.S. (2010). Data Mining with Differential Privacy.493-502.

[9] Chris Clifton, (2001). Privacy Preserving Distributed Data Mining.1-10.

[10] Clifton C.C, & Marks, C.D. (1996). Security and privacy Implications of Data Mining. *In proceedings of the 1996 ACM SIGMOD.*