Application of Data Mining Algorithm in Network Security Detections

Yukun Yang *

University of Massachusetts Amherst, Amherst, United States

* Corresponding author: yukyang@umass.edu

Abstract. In today's world, it is important to recognize online risks. The article is written to let more and more people able to recognize the threats on the network while surfing the Internet. This article is about the definition of data mining and some basic properties. In this paper, the use of data mining technology to prevent network security is examined along with current Internet network security risks. This paper describes several algorithms, including C4.5 and KNN. In addition, this paper describes the network intrusion detection system and explains how data mining methods have been applied to this system. The purpose of this paper is to let more people fully understand and recognize data mining and to be able to use data mining as a method in other fields. In addition, this paper introduces several network security detection methods to describe how data mining can maintain network security.

Keywords: Data mining, Internet security, threats on the network.

1. Introduction

With the development of computers and networks, data mining has gradually become an essential tool for mastering rain management data. Data mining is a new technique that focuses on finding hidden information in a database by collecting data and then building appropriate models [1]. Data mining can be understood as combining machine learning and data statistics [1].

Nowadays, data mining is very widely used in various industries. Due to the rapid development of information technology, more and more people are using the Internet for information transfer and communication. Many companies rely heavily on the web to connect with other companies and users. Some agencies and departments also use computers for confidential document storage [2]. At the same time, the number of hackers is gradually increasing. Hackers are targeting the Internet with increasing frequency. By breaking into networks, hackers steal a wide variety of data to distribute and trade. In order to protect the Internet and avoid the leakage of confidential data, the network's security measures become an important task. Nowadays, through data mining, people can effectively prevent attacks and threats in the network.

This paper focuses on some data mining algorithms introduces a network security detection system and illustrates the application of data mining methods in it.

2. Functionalities of Data Mining

Data mining refers to the rational collection and categorization of data within a specific range and the calculating of the laws between the data. These laws are often not obvious but are hidden in a large amount of data and are not easy to find. Many steps need to be taken to complete data mining. The most important steps are collecting and categorizing data, building and testing models, deriving data, and making predictions [3]. Below is a flowchart of the main processes related to data mining method.

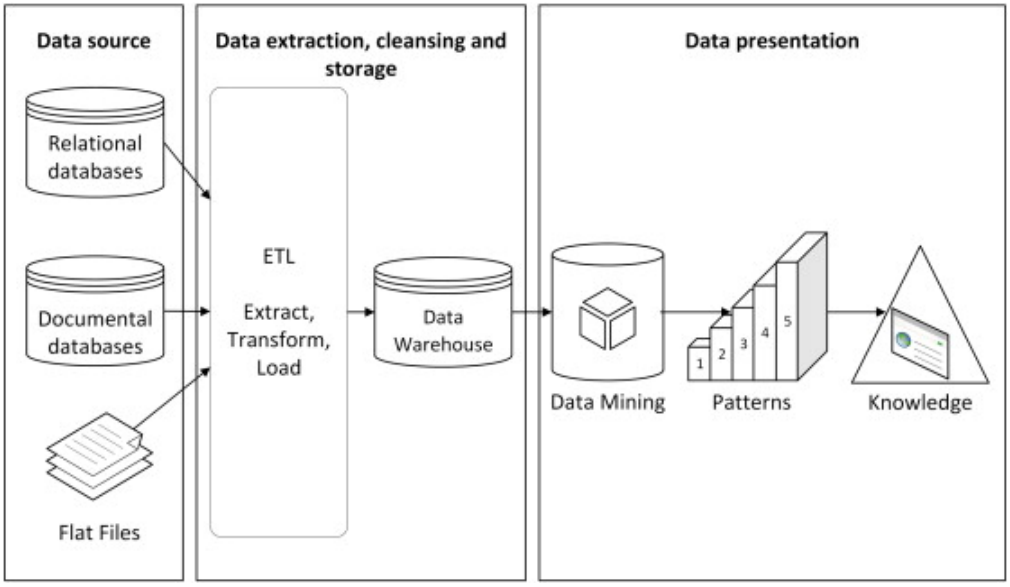

Figure 1. Data mining process [4].

It is crucial to categorize data properly. There are many ways to get data. User data is stored in databases by numerous websites and programs. The data can be obtained by making calls to existing databases. In addition, data can also be obtained through several detection devices (e.g., sensor networks) and algorithms. The sample size of the data should be large enough. Otherwise, it will be challenging to analyze valid information from it. After collecting the required data, we can see from Figure 1 that there are many ways to organize the collected data, such as using relational databases and flat files. When all the data has been organized, it is time for the next progress.

After the data categorization is done, the researcher needs to choose the appropriate algorithm to make predictions from the data that has been obtained. There are many algorithms for data classification today. One of the most widely used ones is the C4.5 algorithm. C4.5 is an algorithm using a decision tree to generate data [5]. A decision tree is like a flowchart tree structure. Each node in the decision tree has at least two child nodes [6]. Arcs will connect each node and their child nodes. Different test results will be noted on each arc between nodes. Another algorithm is called K-Nearest Neighbor (KNN). KNN is a widely used algorithm based on nearest neighbor classification [6]. The main idea of the Nearest neighbor algorithm is to find the nearest point in the data. The KNN algorithm has a higher correctness rate, but it has a very long computing time (O(n2)) [5]. Therefore, many algorithms are modified algorithms based on KNN, such as Wavelet Based K-Nearest Neighbor Partial Distance Search (WKPDS) algorithm and Equal-Average Nearest Neighbor Search (ENNS) algorithm [6]. Bayesian networks are also widely used in data mining. A Bayesian network is a model that describes the probabilities between variables [7]. Bayesian networks can better handle incomplete data sets [7]. Moreover, it can be easily combined with other models, making the data processing more efficient. A researcher can construct a model based on a problem-specific scenario and record data into the model to make predictions [3]. The researcher can also use regression analysis to make predictions. After the data has been analyzed and predicted with a suitable model, some generalization and representation of the results obtained can be done.

Data mining is now widely used in various fields. For example, in business, data mining is often used for customer analysis and can predict future markets based on sales. In the medical field, data mining is often used to record patient information and identify details doctors tend to overlook in large amounts of data.

3. Threats in Internet

As the internet becomes more and more developed, the variety of cybersecurity threats increases rapidly. One of the most common threats is computer viruses. A computer virus is a code that copies

itself to other programs or files on the computer. Computer viruses usually destroy programs and data and may also hold users to ransom. One of the computer viruses that can be harmful is the Trojan horse virus. A Trojan horse disguises itself as a legitimate program or file [8]. When the user installs the file, the Trojan allows the Trojan's developers to manipulate the computer remotely [8]. Through it, the developer can steal the data stored in the user's computer, steal network accounts and passwords, and even shut down the computer at will [8].

There is another threat to computers that comes from hackers. Hackers can access content that is otherwise not allowed to be accessed through some computer vulnerabilities and thus steal some confidential information. After successfully stealing the data from the computer, they will sell the data to other criminals or engage in other illegal activities.

4. Applications in Internet Security

By analyzing and summarizing patterns through data mining, people can have more protection against a wide range of cybersecurity threats. For example, data mining methods can be used for intrusion detection of network databases. Intrusion detection is used to detect unauthorized users of computers and abusive users [9]. Intrusion detection can be categorized into two main groups: Misuse Detection (SD) and Anomaly-based Detection (AD) [10].

Signature-based detection identifies threats by comparing unknown information with captured attacks or threat events. The advantage of this method is that the detection accuracy is high, but it is difficult to detect attacks and threats that have not been recorded [9]. Anomaly-based Detection (AD) is detection through configuration files. It is normal if the network operates according to the intended behavior. The opposite triggers Anomaly-based Detection. This detection is highly accurate and detects new network security threats. However, this detection is computationally expensive because the engine must understand the goals of different users [11]. Moreover, if the malicious behavior is determined to match the predetermined, it will not be easily detected again.

As network architectures become more complex, the data for cybersecurity analytics is growing exponentially. The content of this data is also becoming more diverse [12]. Moreover, the speed of communication and exchange between network data is gradually increasing [12]. These factors make traditional network security detection increasingly tricky. They are using data mining methods that result in more complete data storage and application than traditional network security testing. Also, data mining can visualize and analyze the resulting data. These can make the efficiency of network security detection higher. Data mining can extract patterns from large amounts of data that are not obvious and unknown in advance. Data mining can be utilized to find patterns in many captured attacks and threats that can help detect intrusion. People can use data mining to build intrusion detection models to classify and analyze data from various sources. While using data mining for analysis, data can also be stored using cloud computing technology. This allows the models of data mining to be self-optimized [32]. After using data mining, the types of attacks can be classified into three categories: Masquerading, Penetration, and Intrusion Attempts. Masquerading is an attack by stealing accounts or passwords and can delete more system records. Intrusion Attempt is an attack by guessing the password and causing the operation to fail.

After discovering the pattern of attack types, the pattern and clues can be used to prevent and avoid future intrusions. First of all, people should strengthen their safety awareness. In the process of using computers, attention should be paid to the maintenance of personal privacy. Do not easily browse unfamiliar links; beware of pop-up windows and automatically downloaded software from browsers. For enterprises, they should pay more attention to the security of secrets stored in computers. Secondly, the professional level of relevant personnel should be improved. Many people who use computers do not know much about computers, and it is difficult for them to recognize the authenticity of a large amount of information on the Internet. The relevant departments should actively popularize knowledge about network security. Finally, computer users should actively use

firewalls [11]. A firewall is a computer protection that can isolate some unsafe network elements. The invasion of computer viruses and Trojan horses can be effectively reduced through firewalls.

5. Conclusion

Data mining is a beneficial computer technology. A complete data mining method consists of collecting and categorizing data, building predictive models based on the kind of data, testing and getting results, and discovering non-obvious patterns from the results.

Maintaining network security is necessary in order to have a good network environment. With the increasing types of data and threatening computer programs and attacks in the network, data mining is necessary to protect the network. Data mining is more accurate in processing the data as compared to other methods. Through data mining, one can make network security detection more accurate. Moreover, the data that has been stored is less likely to be lost. Data mining is anticipated to become more effective at identifying network dangers and attacks in the future since ensuring that users use computers safely has long been a concern for all of humanity.

References

[1] Coenen F. Data mining: past, present and future. The Knowledge Engineering Review, 2011, 26 (1): 25 - 29.

[2] Ahmad B, Jian W, Anwar Ali Z. Role of machine learning and data mining in internet security: standing state with future directions. Journal of Computer Networks and Communications, 2018, 2018.

[3] Jassim M, Abdulwahid S. Data Mining preparation: Process, Techniques and Major Issues in Data Analysis. IOP Conference Series: Materials Science and Engineering, 2021, 1090. 012053.

[4] Siguenza-Guzman L, Saquicela V, Avila-Ordóñez E, et al. Literature Review of Data Mining Applications in Academic Libraries. The Journal of Academic Librarianship, 2015, 41 (4), 499 - 510.

[5] Reddy N C S, Prasad K S, Mounika A. Classification algorithms on datamining: a study. International Journal of Computational Intelligence Research, 2017, 13 (8): 2135 - 2142.

[6] Chen F, Deng P, Wan J, et al. Data mining for the internet of things: literature review and challenges. International Journal of Distributed Sensor Networks, 2015, 11 (8): 431047.

[7] Gudipati V K, Vetwal A, Kumar V, et al. Detection of Trojan Horses by the analysis of system behavior and data packets. 2015 Long Island Systems, Applications and Technology, Farmingdale, NY, USA, 2015, 1 - 4.

[8] Britannica. The Editors of Encyclopaedia. Encyclopedia Britannica, 2022.

[9] Li W. The network database intrusion detection system based on data mining. Computer knowledge and technology, 2018, 14 (36): 7 - 8.

[10] Liao H J, Lin C H R, Lin Y C, et al. Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 2013, 36 (1): 16 - 24.

[11] Jyothsna V, Prasad R, Prasad K M. A review of anomaly-based intrusion detection systems. International Journal of Computer Applications, 2011, 28 (7): 26 - 35.

[12] Liying Y. Application of big data technology in Network security analysis. Electronic Technology and Software Engineering, 2022 (11): 50 - 53.

[13] Lijuan Y, Feng S. Application of data mining technology in computer network intrusion detection. Integrated Circuit Applications, 2023, 40 (07).