Advances in Social Science, Education and Humanities Research, volume 543
Proceedings of the 2021 6th International Conference on Social Sciences and Economic Development (ICSSED 2021)

Application of Time Series and Clustering in Research of Information Vulnerabilities

Jing Nan[1,*] Ruiying Jin[2]

[1] *School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100000, China*
[2] *School of Economics and Management, Beijing Jiaotong University, Beijing, 100000, China*
Email: 18721057@bjtu.edu.cn, 18711052@bjtu.edu.cn

ABSTRACT
With the advent of the era of big data, computer networks have penetrated into every aspect of people's life and work. While people enjoy the convenience brought by the Internet, they also face the threat of network security vulnerabilities, and it is urgent to strengthen network security. China National Information Security Vulnerability Database classifies information security vulnerabilities into 26 categories, such as configuration errors and SQL injection. With the increasing complexity of information security in the past two years, the traditional vulnerability classification standard has been subjected to new tests. In this paper, we use crawler technology to obtain vulnerability data from China National Information Security Vulnerability Database from 2004 to September 2020, time series analysis to predict the number of vulnerabilities, and use text analysis and clustering technology to re-cluster the vulnerabilities and derive new vulnerability classification levels and criteria.

Keywords: *information security, clustering, text analysis, time series analysis*

1. INTRODUCTION

With the development of computer technology and the improvement of information technology, the emergence of various Internet applications has brought great convenience to people's lives. Online shopping, online bill payment and other functions have changed the traditional shopping mode, and teleconferencing software such as Zoom and Teams have made telecommuting possible. Information technology has become an important force for economic development, profoundly changing every aspect of people's lives. The following chart shows the development of China's big data industry scale in recent years.

The development of informatization and technology has made the connection between cyberspace and real life even closer. The network penetrates into the real life and constitutes an open and complex system. Information security is not just the reliability and stability of the network, but has a more realistic interpretation and meaning. Information security is not only a key factor affecting the healthy development of cyberspace, but also an important part of social stability and healthy development.

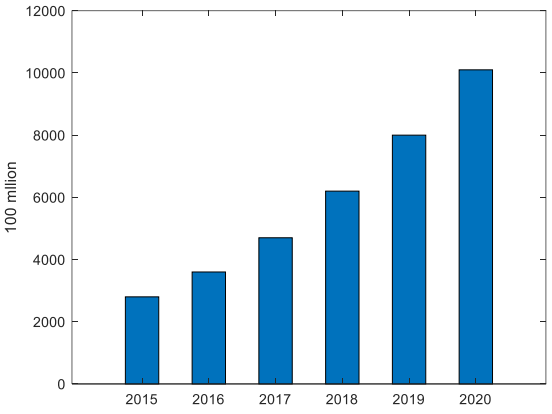

Figure 1 Big data industry scale.

1.1. Vulnerabilities

A vulnerability is a weakness or flaw in a system, the susceptibility of a system to a specific threat attack or dangerous event. Vulnerabilities often arise from errors created in the design or coding of software or operating systems. The presence of a vulnerability can easily lead to hacking and virus residency, resulting in data loss and tampering, user information and privacy disclosure.

1.2. Data Source

China National Information Security Vulnerability Database (CNNVD) is an official vulnerability database established and maintained by China Information Security Testing and Evaluation Center, which is mainly responsible for vulnerability analysis and risk assessment. In the CNNVD vulnerability repository, each vulnerability uniquely corresponds to a CNNVD number. The record form of vulnerability information contains 14 fields, including CNNVD number, hazard level, vulnerability profile, etc. According to the hazard level, vulnerabilities can be classified as low, medium, high, and ultra-risk.

The specification adopted by CNNVD divides information security vulnerabilities into 5 hierarchical relationships totaling 26 types. They are lack of message,code, configuration, resource management errors, improper input validation, numeric errors, information exposure, security features, race condition, buffer errors, injection, path traversal, link following, authentication issues, insufficient verification of data authenticity, credentials management, permissions and access controls, cryptographic issues, format string, command injection, cross scripting, code injection, SQL injection, cross-site request forgery, improper access control, OS command injection.

2. TIME SERIES ANALYSIS

In order to explore the relationship between the number of information security vulnerabilities and time, and to predict the future number, this paper uses the time series analysis method to establish an Auto-regressive moving average model (ARMA) model to analyze the development process and development pattern by observing and aggregating the total number of vulnerabilities found and registered each year between 2004 and 2019.

2.1. Time Series Model Background

Time series analysis is a classical concept in statistical analysis. By observing and evaluating a realistic and real chronological set of real data, and using curve fitting methods to establish an objective analysis and description of the system, and to predict the future values of that time series.

2.2. Model design and construction

This paper uses information security vulnerability data from the CNNVD repository between 2004 and 2019. Statistics by year are analyzed and plotted using matlab, and the results are shown below.

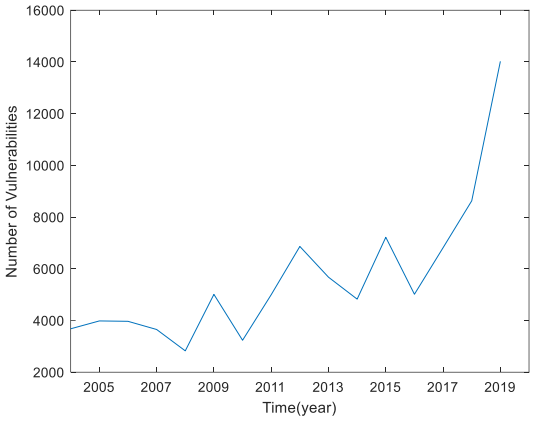

Figure 2 Number of vulnerabilities.

From the graph, it can be seen that the number of information security vulnerabilities announced each year fluctuates, but the trend shows that the number of vulnerabilities is generally on an upward trend. The upward trend with time confirms the scientificity and reasonableness of the time series from the side.

First, before building the model, the data were tested for smoothness, and the Augmented Dickey-Fuller test (ADF) unit root test was used for this modeling. the smoothness test algorithm proposed by Kwiatkowski, Plillips, Schmidt and Shin in 1992. In this paper, the KPSS algorithm is used to perform the test again. The results of the two tests on data smoothness are as follows.

Table 1 Result of ADF and KPSS test

Type	ADF Test	KPSS Test
Value	0	0

According to the test results in the table, it can be seen that the data passed the ADF Test and KPSS Test, satisfying the requirements for the use of time series, and there is no need to use differencing. So in this research, the ARMA model is used for the prediction of the number of future vulnerabilities.

2.3. Determining ARMA model parameters

For the determination of parameters, the classical auto-correlation function and partial auto-correlation function were used in this study, and the ARMA model parameters were determined by the function results.

2.3.1. Auto-correlation Function

Auto-correlation is the comparison of an auto-regressive ordered sequence of random variables with itself. The auto-correlation function reflects the correlation between the values of the same sequence in different time series. The equation for the auto-correlation function is as follows.sentence, as in

$$ACF(k) = \rho_\kappa = \frac{Cov(y_t, y_{t-k})}{Var(y_t)} \quad (1)$$

Based on the image trend of the auto-correlation function, the parameters of the ARMA model are determined by determining whether the function image is a truncated or trailing state. The parameters of the moving average model are determined by the auto-correlation function.

Using Matlab for plotting, the function image goes as follows.

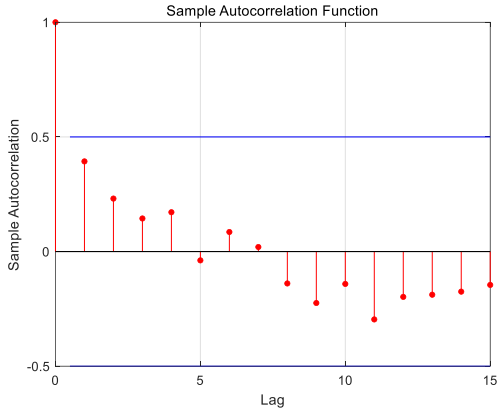

Figure 3 Sample auto-correlation function.

According to the image, it can be seen that the value of the auto-correlation function drops abruptly in the interval from 0 to 1, after which it remains more stable and therefore the function presents a truncated state.

2.3.2. Partial Auto-correlation Function

The partial auto-correlation function is a method for characterizing the structure of stochastic processes. The partial auto-correlation function PACF describes the linear correlation between the time series observations expected to be past observations given intermediate observations. Using the partial auto-correlation function, the coefficients required for the ARMA model can be determined based on the trend of the function image. The image of the partial auto-correlation function for information security vulnerabilities is as follows.

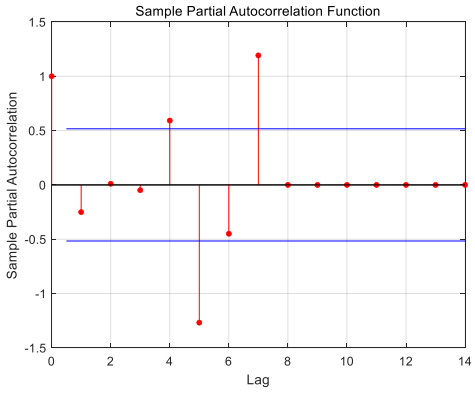

Figure 4 Sample partial auto-correlation function.

2.4. Model Predictions

The ARMA model was selected by testing the stability of the data. Then, in this paper, the values of each parameter of the model were determined by the classical auto-correlation function and partial auto-correlation function. The design of the model is now complete. Then the model is built using the built-in functions of Matlab and the model is used to make predictions. Due to the small amount of data and the change in the number of information security vulnerabilities is affected by many factors, therefore, this paper only predicts the number of vulnerabilities for the next three years. In addition, this paper set a 95% confidence interval, the specific prediction results are shown in the figure below.

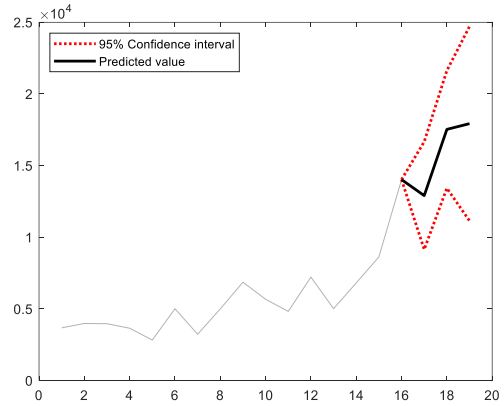

Figure 5 Result of predict.

The solid black line in the figure shows the predicted results. The area between the two red dashed lines is the 95% confidence interval. From the prediction results, under the premise that the development rate of network conditions and computer technology remains stable, the number of information security vulnerabilities included in CNNVD each year in the next three years has a higher probability of showing an upward trend, and the prediction results based on the time series are 12910, 17530, 17930.

3. CLUSTERING

With the development of computer networks and information technology in recent years, the information security posture has become increasingly complex. In order to deal with vulnerabilities more efficiently, CNNVD has taken many measures, one of which is to classify vulnerabilities. 26 types of vulnerabilities are classified by CNNVD according to their characteristics and causes. These 26 types have been mentioned above. Vulnerabilities of the same type often have the same or similar handling methods. Therefore, when programmers find vulnerabilities in software or systems, they can then take possible approaches to quickly resolve the vulnerabilities according to the category to which the vulnerabilities belong. But in today's rapidly

evolving computer technology, new types of vulnerabilities are likely to emerge. Therefore, in order to evaluate whether the 26 classification methods can meet the current requirements for the exploitation of vulnerability information, this paper adopts a clustering analysis algorithm. By clustering more than 140,000 vulnerability profiles recorded in CNNVD, new vulnerability classification methods are explored.

Considering the large number of vulnerabilities, the clustering analysis model is built using python. After comprehensive consideration, the k-means algorithm was adopted for this experiment. 26 groupings are more reasonable, so this paper calculates the degree of aggregation by presetting the groupings to be between 22 and 30, and then using the k-means algorithm.

Since the vulnerability profile column in CNNVD is in Chinese, the Chinese text is split using jieba splitting and common auxiliary words are removed in the data preprocessing stage. Then tf-idf values are calculated. Then k-means is used for clustering. The clustering results were evaluated using the error sum of squared errors (SSE). The error sum of squares formula is as follows.

$$SSE = \sum_{k=1}^{K} \sum_{p \in C_k} \left| p - m_k \right|^2 \quad (2)$$

Since the data exceeds 140,000 items, the calculation workload is too large if all the data is used, so this paper adopts random sampling within the interval and uses 5,000 items of data for clustering calculation, and the final calculation results are as follows.

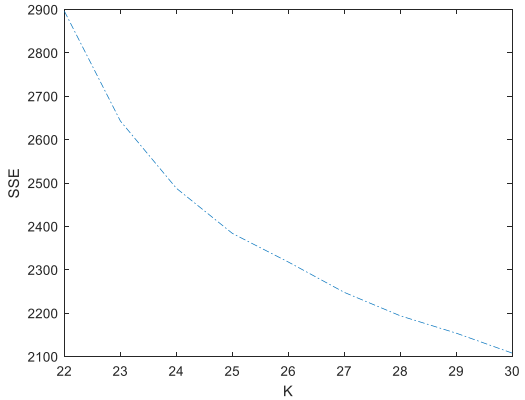

Figure 6 Result of SSE.

As can be seen from the figure, the curve tends to level off between 26 and 28. Therefore, it can be seen from the clustering results of the vulnerability profiles that 26, 27, and 28 can all be considered as reasonable numbers of groupings.

4. CONCLUSION

This paper uses over 140,000 vulnerability messages recorded by CNNVD between 2004 and 2019. An ARMA model was developed using time series analysis and the total number of vulnerabilities was predicted for the next three years. Given the overall upward trend in the number of vulnerabilities and the increasing complexity of vulnerabilities, an evaluation of the current vulnerability classification criteria is conducted. It was explored that 26 to 28 is a reasonable number of vulnerability groups.

REFERENCES

[1] Vergetis Vangelis, et al. “ Assessing drug development risk using Big Data and Machine Learning,” Cancer research ,2020.

[2] Junaid Akram, Luo Ping.(2020).“How to build a vulnerability benchmark to overcome cyber security attacks,” IET Information Security(1), 2020.

[3] Wang Lingyu, Lyu Baolei,and Bai Yuqi.“Global aerosol vertical structure analysis by clustering gridded CALIOP aerosol profiles with fuzzy k-means,” The Science of the total environment 761, 2020.

[4] Annushree Bablani, et al. “A multi stage EEG data classification using k-means and feed forward neural network,” Clinical Epidemiology and Global Health 8.3, 2020.