Journal of Digital Information Management

JDIM

DLINE JOURNALS

ISSN 0972-7272

Identifying Common Cause Failures using Score Data Mining

Yonghui Ma
Xi'an Translation Institute
Xi'an, Shaanxi, China
15395564841@163.com

ABSTRACT: *In this study, we employed data mining to accurately evaluate the failure rate of secure computers, providing valuable data information for our decision-making layer. This technique is beneficial not only for our decision-making but also for the long-term operation of our systems. Through in-depth analysis, we discovered inherent connections among various failure events and their mutual impacts. These findings contribute to a deeper understanding of common cause failures in secure computer systems and prepare for enhancing their security. Establishing a robust information system is essential to meet the increasing demands, especially in the complex internet environment where secure digital computer systems face ever-growing challenges.*

Subject Categories and Descriptors: [H.2.8 Database Applications]; Data mining [K.6.5 Security and Protection] [C.2 COMPUTER-COMMUNICATION NETWORKS]: Data communications;

General Terms: Data Mining, Data Failures, Data Errors

Received: 29 March 2024, Revised 29 August 2024, Accepted 20 September 2024

Keywords: Data Mining, Secure Computers, Common Cause Failure, Decision Tree Algorithm

Review Metrics: 0/6; Review Score: 4.58; Inter-reviewer Consistency: 89.5%

DOI: https://10.6025/jdim/2024/22/4/137-142

1. Introduction

With the rapid progress of technology, secure computer operating systems have become increasingly prominent [1]. They are the core of the business world and the foundation of various industries such as politics, military, and medicine. However, due to their complex architecture and continuous risk factors, they still face severe challenges, such as hardware damage, software crashes, and human operational failures [2]. With the advancement of science and technology, more researchers are dedicated to developing more advanced, secure computer operating systems to ensure their stability and reliability [3]. Data mining is an important component of these efforts as it can effectively reveal complex information and better assess different risks. "Common cause failure" describes a set of complex, closely related failures that significantly negatively impact overall operations, reducing the overall efficiency. To improve the operational status effectively, we need to conduct in-depth research on all failures to identify their root causes, make targeted improvements, and enhance overall stability and reliability. This study aims to develop a new approach to calculate the common cause failure score and evaluate the security performance by comparing the results of different systems. For this purpose, we collected failure event data from secure

computer systems. We used data mining techniques, such as association rule mining and decision tree mining, to explore the characteristics of these failure events [4], to achieve more accurate results.

Through in-depth research, we discovered complex relationships among various failure events and established a decision tree to describe these relationships. To better assess the impact of failure events on system security, we proposed a new calculation method, the common cause failure score. Through comparative analysis, we can identify key failure events in the system and take effective measures to improve the security performance for optimal operation. This study enables a profound understanding of common cause failures in secure computer systems, providing effective decision references for system managers. By comparing the common cause failure rates of different systems, we can discover differences in their security performance, offering strong guidance for selecting systems with better performance [5]. Protecting system stability and reliability is crucial, especially in today's complex network environment, as secure computer systems face increasing threats and challenges. By accurately evaluating and comparing common cause failures of secure computers through data mining techniques [6], their stability and reliability are greatly improved, making them a new technological means. Furthermore, this technology can help companies more accurately identify hidden dangers and take corresponding measures to prevent these risks, further enhancing security.

2. Related Work

In recent years, the research field of secure computer systems has made considerable progress, from common cause failure analysis to system performance evaluation, employing various methods and techniques to significantly enhance the stability and reliability of secure computer systems. This section will delve into some important content related to FTA. FTA (data mining, data integration, data modeling) is widely applied in security reliability assessment, effectively identifying system problems and providing valuable information for decision- makers [8]. Although traditional fault tree analysis methods can explore system structural characteristics from a single perspective [7], they fail to capture the complexity of various failure events. Therefore, more powerful measures are needed to delve into these complexities, such as using association analysis to better analyze the mutual influence of various failure events and protect network stability. A new study utilized association rule mining to explore the connection between hardware failures and software defects in secure computer systems. The results show that using this technique can discover their connections [9] and reveal their internal linkage, exposing their common causes. Neural network technology has been widely applied in various fields, and it is known for its ability to learn and process complex information [10], which is especially suitable for dealing with security issues. An interesting experiment demonstrated that it can help us identify and address various hardware issues, thus enhancing our security. Through in-depth analysis of neural networks, we found that it can quickly detect hidden hardware issues, providing users with precise diagnostics and solutions [11,12]. Through multiple-factor analysis, we can better assess system performance and security. This method combines multiple factors to more accurately predict the occurrence of failure events and evaluate the impact of each factor on system security performance. Through multiple-factor analysis, a study found that hardware, software, and human operational errors adversely affect the performance of secure computer systems. Although these studies have achieved certain results, challenges still require further research to address. While some methods can effectively detect failures in the system, they lack the correlation between these events, thus failing to comprehensively assess the security performance of the system [13]. Additionally, these methods require substantial human resources and expertise, limiting their feasibility in practical applications. This study aims to utilize data mining techniques to construct a new common cause failure score model to compare and analyze the security performance of different systems and explore the correlation and influencing factors among failure events, thus achieving a better grasp and improvement of secure computer system stability and reliability.

3. Design of Classification Algorithms

This paper will introduce several common classification algorithms and discuss their applications in secure computer systems to provide strong support for enhancing security performance. When building decision trees, attribute selection is essential as it determines the size of the decision tree and the final classification result [14,15]. Therefore, we proposed a new algorithm that introduces a balance coefficient to reduce the information gain of attribute A, effectively selecting and optimizing the testing attribute. The formula for this algorithm is as follows:

$$BC(A) = \frac{\sum_{i=1}^{n} | S_{i1} - S_{i2} |}{n} \quad (1)$$

BC(A) represents the balance (corrected) coefficient, and n represents the total number of training samples in the set S. The Naive Bayes algorithm aims to determine the inductive relationships of a set of data and the relationships between various factors in these relationships, thereby identifying which factors play an important role in the inductive relationship of this data. The Naive Bayes algorithm is a useful method for identifying and predicting security issues. By comparing different data sets, it can effectively identify and assign certain failure information

to a specific security risk level. This method is also widely used in document analysis and recognition. By calculating the premises and assumptions of the samples, we can estimate which specific distribution they belong to based on the Bayes formula. This formula indicates that if the premises and assumptions of a distribution are correct, it belongs to the characteristics of that distribution.

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)} \quad (2)$$

Let *T* = {(*x1, y1), (x2, y2*), ..., (*xK, yK*)} be the given data set, where the input samples are {*x1, x2, ..., xK*}, and the output classes are {*y1, y2, ..., yK*}, with K as the number of samples. Each input sample consists of n dimensional features, i.e., {{*x(1)1, x(2)1, ..., x(n)1*}, {*x(1)2, x(2)2*, ..., *x(n)2*}, ..., {*x(1)K, x(2)K, ..., x(n)K*}}. Let X and *Y* be random vectors, where X belongs to the feature space, and *Y* belongs to the class space, and we use Y = cm to represent the *m-th* class in the class space, where *m = 1*, *2*, ..., *M, and M* is the total number of classes. The fundamental concept of the Naive Bayes algorithm is to view X and Y as a set of random vectors, where *X* represents a specific set of probabilities. *Y* represents a set of classes, with the size of each class being calculated through probabilities *m*, *2*, ..., M, M, etc., to determine the class of the sample. When the range of x(j) is from 1 to m, i.e., from Sj to n, i.e., from *M n j=1Sj* to *M n j=1Sj,* such settings may lead to the number of parameters exceeding expectations, increasing computational difficulties. Therefore, efforts should be made to minimize such occurrences and improve the efficiency of the settings. Hence, when using the Naive Bayes algorithm, we can control the possibilities through a specific assumption, which may take the form of:

$$\prod_{i=1}^{n} P\left(X^{(j)} = x^{(j)} \mid Y = c_m\right) \quad (3)$$

Although this assumption may affect the algorithm's classification accuracy, it helps to reduce the computational complexity, making the model more refined and easier to apply in various scenarios.

4. Experimental Design and Analysis

In data mining, experimental design and analysis are vital for computing the common cause failure scores and conducting comparative analysis of security computing systems. Through carefully designed experiments and detailed data analysis, we can obtain more accurate results and better infer the root causes of issues. Through analysis, we found that the HI (Health Index) evaluation model of the control center can accurately reflect the health status of the organization [16-18]. We use random fuzzy distributions to represent this information in order to understand their distribution better. By analyzing the HI data from the past few years, we used MATLAB for least squares fitting to derive the HI distribution functions for each unit, as shown in Figure 1.

Through carefully designed technical schemes, we can accurately assess the health status of the security host unit and use a life prediction model to predict its remaining life. From Figure 1, it can be observed that the distribution of HI for the constituent units exhibits distinct exponential distribution characteristics. Data preprocessing can eliminate unnecessary information from the raw data, such as missing, abnormal, and duplicate data. Additionally, we can employ statistical methods, transformers, and dimensionality reduction

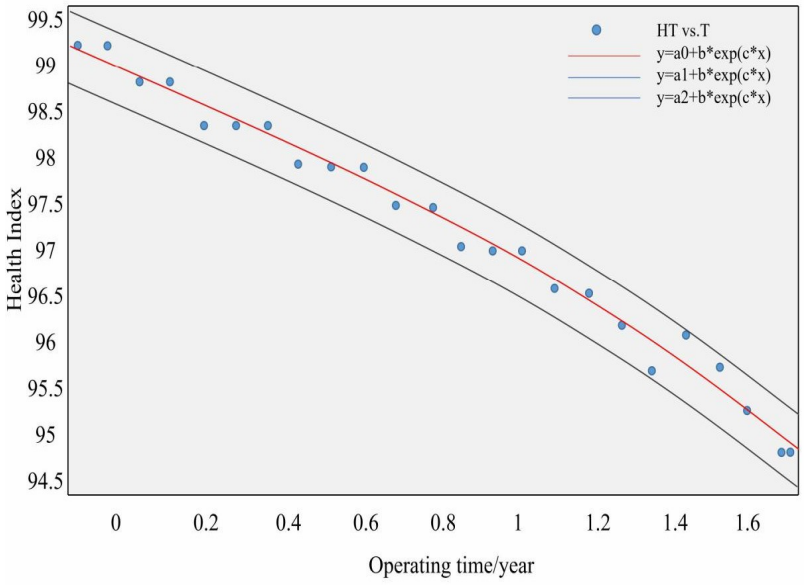

Figure 1. Changes in Health Index (HI) of Control Center Components over Time

algorithms to unearth valuable information and better meet our research needs. Before conducting computer state monitoring, it is necessary to select the most appropriate information from multiple parameters, including the clock, power supply, ambient temperature, instantaneous pressure (0 denotes normal, 1 denotes instantaneous pressure alarm), last runtime (0), current status (0), and final result. By utilizing the SPSS data file 1, we can access abundant data and apply data mining techniques to obtain more accurate results, as shown in Figure 2.

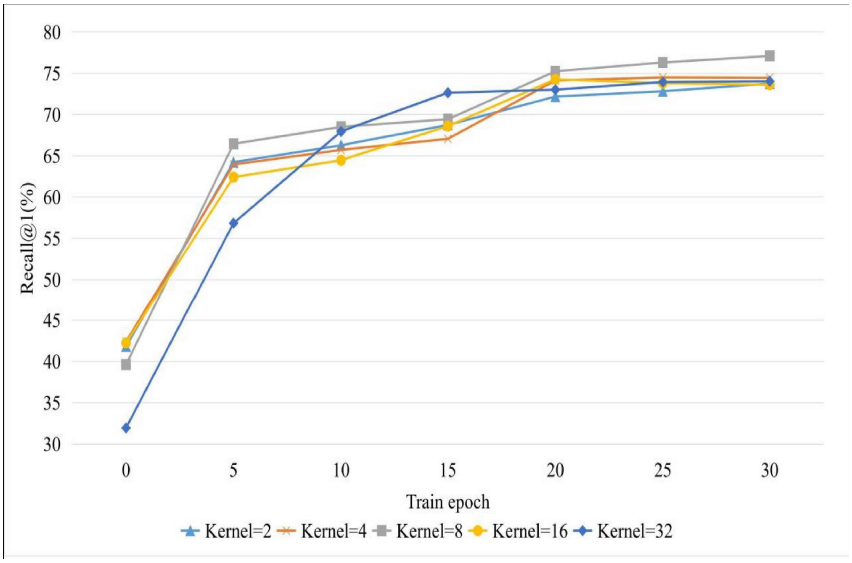

Figure 2 . Co-Failure Score of Decision Tree Algorithm

By introducing a large number of input and output nodes, multiple parameters can be generated, including the instantaneous changes in temperature, power, and other parameters, as well as the visibility of these changes (if the value of the previous parameter and the current parameter is opposite, i.e., peak or valley, then the value is true), and the stability of the parameters (initially in a steady state, but transitioning to a changing stage when two continuous changes are detected). The system is switched to a stable mode if no power fluctuations occur within 5 periods and no adjustments are needed. Additionally, statistical analysis should be performed on the temperature and related spectrum data for the most recent 5 periods. The first spectrum data should be the final reference to avoid excessive errors. After screening, we are left with the latest information containing the dates, status, data, pressure reports, functionalities within the most recent 5 periods, and the latest data for temperature and functionalities within these dates.

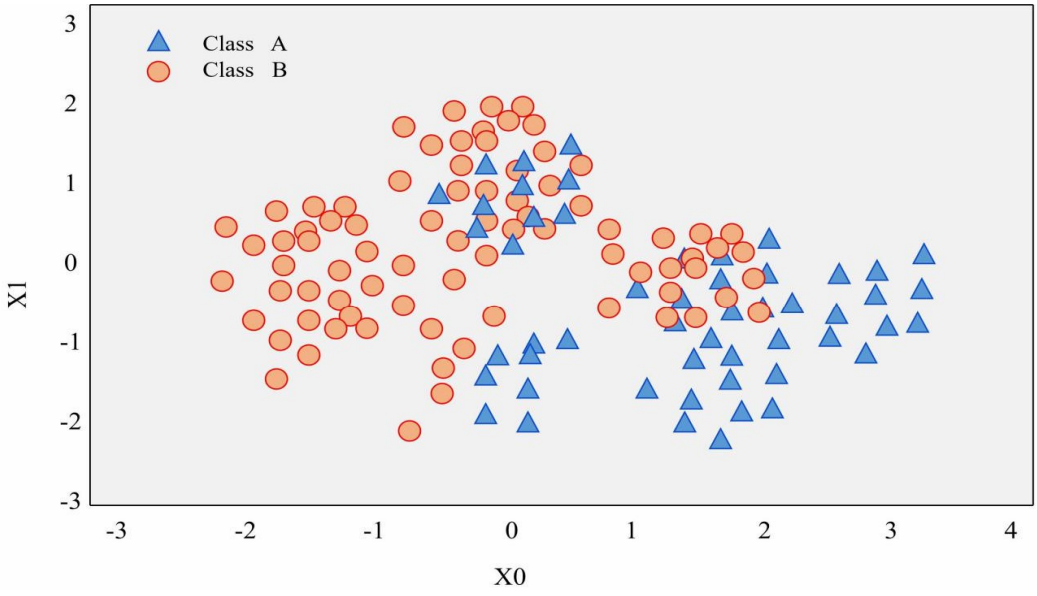

Figure 3. Naive Bayes Co-Failure Data Chart

Using neural networks and the C5.0 algorithm, we can effectively transform complex information into more precise, reliable, and controllable results, enabling us to conduct effective research on complex information sets. To better meet the requirements of different algorithms, comprehensive and efficient information transformation is needed for all algorithms. Figure 3 shows the prepared steps. We used graphical tools such as ROC curves, confusion matrices, and bar charts to present our research results better. Furthermore, we will investigate these results further, including their advantages and disadvantages, the impact of key information, and their similarities. The effectiveness of the experiments will be verified through various effective data mining skills, such as variance analysis, t-tests, and correlation studies. To obtain more accurate conclusions, the sample size should be considered fully, and all factors that may affect the findings should be strictly controlled to achieve better accuracy. When taking any measures to collect or process information, attention should be paid to the integrity, sensitivity, and completeness, and all relevant laws and regulations should be strictly adhered to. Through careful experimental design and in-depth research on a large amount of information, we can better understand the causes of computer failures and provide better guidance for system operation and risk control.

5. Conclusions

Using data mining techniques, we can accurately assess the failure rate of systems and identify potential hazards. This method not only ensures the system's stable operation but also helps system managers identify the causes of failures and take appropriate measures to prevent them. This article has explored several commonly used ranking methods, including decision trees, naive Bayes, support vector machines, and random forests, and how they are widely applied in secure computing. Through detailed experimental research and data analysis, we have determined that each method has its unique advantages in better meeting various complex situations for evaluating security. We can find a more accurate solution using the decision tree algorithm. In contrast, the naive Bayes algorithm can handle complex multidimensional data more effectively, and the support vector machine algorithm is better suited to solve complex linear and nonlinear problems. The random forest algorithm is an effective decision tree model that combines multiple decision tree models effectively and improves the model's performance in a targeted manner, making it more refined and effective in predictions. Moreover, it can help better identify common failures and assess the model's effectiveness. Through in-depth research, we can accurately measure the stability and security of a system, enabling us to design targeted maintenance and optimization plans. When these indicators reach a specific level, it indicates that the system has been influenced to some extent, prompting managers to take decisive measures to address the issue. Using data mining techniques, we can calculate and compare the causes of failures in secure computing, providing valuable assistance for system maintenance and risk management. Furthermore, we can obtain accurate results through carefully designed experiments and practical data analysis, providing valuable recommendations and improvement plans for system managers. Implementing these measures will significantly enhance the reliability and stability of secure computing systems and effectively reduce potential risks and losses.

References

[1] Rao, Z. and Yuan, J. (2021). Data mining and statistics issues of precision and intelligent agriculture based on big data analysis. *Acta Agriculturae Scandinavica, Section B - Soil & Plant Scienc*e, 2021 (4) 1-14.

[2] He, Y., Chu, Y., Song, Y., others. (2022). Analysis of design strategy for energy-efficient buildings based on databases using data mining and statistical metrics approach. *Energy and Buildings*, 258, 111811.

[3] Ngow, Y. T., Goh, S. H. (2021). Diagnostic-driven yield engineering under atypical wafer foundry conditions. Microelectronics Reliability, 1191) 114076. https://doi.org/10.1016/j.microrel.2021.114076

[4] Merani, M. L., Croce, D., Tinnirello, I. (2021). Rings for Privacy: An architecture for large scale privacy-preserving data mining. *IEEE Transactions on Parallel and Distributed Systems*, p. 99, 1-1.

[5] Wang, A., Yu, H. (2022). The construction and empirical analysis of the company's financial early warning model based on data mining algorithms. *Journal of Mathematics*, 2022.

[6] Baykara, M., Abdulrahman, A., Alahmed, A. S. (2022). Classification of network data with machine learning methods for intelligent intrusion detection systems. *In*: 2022 4th International Conference on Advanced Science and Engineering (ICOASE) (p. 77-82). IEEE.

[7] Suo, N., Zhou, Z. (2021). Computer assistance analysis of power grid relay protection based on data mining. *Computer-Aided Design and Applications*, 18 (S4) 61-71.

[8] Zhou, S. (2021). A machine-learning method of predicting vital capacity plateau value for ventilatory pump failure based on data mining. *Healthcare*, 9.

[9] Guo, H. X., Wang, J. R., Peng, G. C., et al. (2021). A data mining-based study on medication rules of Chinese herbs to treat heart failure with preserved ejection fraction. *Computer-Aided Design and Applications*, 28 (9) 8.

[10] Lin, Z., Xiangping, L., Wenzhong, C., and et al.

(2021). Computer aided analysis and control of power system based on data mining technology. *In:* 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC). IEEE.

[11] Makuvaza, A., Jat, D. S., Gamundani, A. M. (2021). Deep neural network (DNN) solution for real-time detection of distributed denial of service (DDoS) attacks in software defined networks (SDNs). *SN Computer Science*, 2, 1-10.

[12] Chiba, Z., Abghour, N., Moussaid, K., et al. (2019). New anomaly network intrusion detection system in cloud environment based on optimized back propagation neural network using improved genetic algorithm. *International Journal of Communication Networks and Information Security*, 11 (1) 61-84.

[13] Cecotti, H., Graser, A. (2010). Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 (3) 433-445.

[14] Wang, Z., Luo, N., Zhou, P. (2020). GuardHealth: Blockchain empowered secure data management and Graph Convolutional Network enabled anomaly detection in smart healthcare. *Journal of Parallel and Distributed Computing*, 142, 1-12.

[15] Abdallah, A. E., Hamdan, M., Gismalla, M. S. M., and others. (2023). Detection of management-frames-based denial-of-service attack in wireless LAN network using artificial neural network. *Sensors*, 23 (5) 2663.

[16] Tuli, S., Basumatary, N., Gill, S. S., et al. (2020). HealthFog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments. *Future Generation Computer Systems*, 104, 187-200.

[17] Riaz, M. S., Qureshi, H. N., Masood, U., and et al. (2022). Deep learning-based framework for multi-fault diagnosis in self-healing cellular networks. *In:* 2022 IEEE Wireless Communications and Networking Con*ference* (WCNC) (p. 746-751). IEEE.

[18] Sunitha, G., Arunachalam, R., Abd-Elnaby, M., et al.. (2022). A comparative analysis of deep neural network architectures for the dynamic diagnosis of COVID-19 based on acoustic cough features. *International Journal of Imaging Systems and Technology*, 32 (5) 1433-1446.