Computer Fraud and Security
ISSN (online): 1873-7056

Data Security Risk Assessment and Control based on Improved Apriori and NSGA-II

Xixiang Zhang[1], Yun Dong[1*], Xuhua Ai[1], Yuan Yin[1], Qi Meng[1], Zhaoli Chen[1], Liyuan Zhang[1], Zhipeng Meng[1], Kaijie Liu[1]

[1]Information Center, China Southern Power Grid Guangxi Power GridCo Ltd, Nan ning, 530015, China.

Abstract:
With the rapid development of information technology, the prominence of net work data security risks is escalating. Accurately evaluating these risks and devising effective control strategies have emerged as urgent imperatives. To address this challenge, this study introduces an enhanced version of the Apriori algorithm to discern correlations within data and perform corresponding risk assessments. Additionally, it employs the elitist non-dominated sorting genetic algorithm for data security risk control. The findings demonstrate that the probability of occurrence of primary risk factors, as calculated using the enhanced Apriori, exceeds 0.7, with approximately 75% confidence in the association between each factor and medium to high risk. These calculations align closely with the ”main network security risks” outlined in the unit’s 2021 network security work summary. Furthermore, the distribution of Pareto optimal solution sets derived from the multi-objective evolutionary algorithm based on decomposition exhibits non-uniformity, whereas those obtained from the elitist non-dominated sorting genetic algorithm manifest diverse and evenly distributed outcomes. Moreover, considering risk state values and control costs, this study adopts a third approach to mitigate security risks concerning application systems and data. This strategy yields a risk level of only 0.386 and incurs a cost consumption of merely 2,404,800 yuan. The proposed data security risk assessment and control strategy demonstrate strong feasibility, effectively enhancing the value of data utilization and delivering practical benefits to enterprises or organizations.

Keywords: Data security risk, Risk quantification, Security threats, Data value, Improved Apriori, NSGA-II

1 INTRODUCTION

As the Internet and information technology advance rapidly, enterprises and organizations are confronted with significant threats to network data security. Incidents such as data leakage, unauthorized access, and malicious attacks not only result in substantial economic losses but also jeopardize the reputation of enterprises and erode customertrust[1]. Presently, mainstream approaches for analyzing Data Security Risk (DSR) in networks encompass qualitative, quantitative, and comprehensive analyses. Qualitative analysis predominantly relies on the expertise and intuition of analysts, assigning numerical values to each indicator through expert discussions on scoring systems, and utilizes a fixed mathematical model to generate results. This method exhibits notable subjectivity and high reliance on human input, often failing to fully encapsulate the risks inherent in the system. Quantitative analysis entails establishing an evaluation system grounded in specific standards, accurately quantifying various indicators in a singular evaluation, and deriving risk values through mathematical modeling. However, this method frequently struggles with quantitatively evaluating system aggregations or large-scale information systems, with resultant reliability concerns. Comprehensive analysis amalgamates both qualitative and quantitative methods, flexibly selecting different combinations based on real-world application scenarios to yield results tailored to enterprises or systems. While this method offers distinct advantages, its implementation is intricate, necessitating the construction of specialized analysis methods based on practical and historical experiences. Although these three methods are suitable for individual systems or small system clusters, they prove inadequate for situations characterized by complex organizational structures and numerous systems[2, 3]. To address such challenges, this study introduces and applies an improved version of the Apriori algorithm for network DSR evaluation and employs an Elitist Non-Dominated Sorting Genetic Algorithm (NSGA-II) for DSR control.

The research encompasses four main components. Firstly, a comprehensive review of Apriori and NSGA-II is conducted. Secondly, the DSR evaluation and control strategy employing the enhanced Apriori and NSGA-II are

outlined. Subsequently, performance validation and application analysis of the proposed method are undertaken. Finally, the findings are summarized, discussed, and future prospects are delineated.

2 RELATED WORK

The Apriori algorithm, commonly utilized in data mining, is employed to extract frequent itemsets and Association Rules (AR). Santoso M H advocated for the use of AR and Apriori to analyze transaction data, aiming to unveil customer purchase patterns. The findings of this study confirmed a recurring trend wherein customers frequently purchased toothpaste and detergent simultaneously, meeting the minimum confidence threshold. Employing this Apriori-based search pattern was anticipated to enhance sales strategies [4]. In response to the spatial and temporal complexity inherent in identifying flexible periodic patterns in time series databases, Hendrasty N et al. introduced an enhanced version of Apriori. By leveraging prior methodologies and hash generation vectors, the algorithm efficiently unearthed various types of flexible periodic patterns while adeptly managing information in hash tables. The outcomes indicated that the proposed approach exhibited reduced time and space requirements when handling extensive datasets compared to conventional methods [5]. Abidin Z et al. proposed a refined iteration of the Apriori algorithm to address the escalating data trading volume within the largest automotive company. Parameters such as minimum support and confidence were predominantly configured, with prior algorithms employed to sift through effective studies. The study's findings validated that this method facilitated the elimination of frequent itemsets that compromised the lowest level of trust, thereby enhancing the accuracy and reliability of data mining endeavors [6]. Styawati S et al. applied the Apriori algorithm within RapidMiner to scrutinize customer purchase patterns gleaned from sales transaction data at Diengva store. Their investigation unveiled that items meeting the minimum support and confidence criteria encompassed eyelashes, eyelash gel, soft lenses, and soft lens water. The robust confidence values associated with these rules could inform inventory stocking decisions and furnish tailored inventory management strategies for stores [7].

The NSGA-II algorithm is commonly employed to tackle optimization problems characterized by multiple conflicting objectives. Zheng W et al. conducted mathematical runtime analyses and determined that NSGA-II exhibited robust performance on fundamental benchmark functions like OneMinMax and LOTZ, provided the population size remained consistently larger than the Pareto frontier size. However, when these sizes equated, NSGA-II encountered challenges in accurately computing the entire Pareto front [8]. In addressing environmental multi-objective planning within reconfigurable manufacturing environments, Khettabi I et al. introduced novel dynamic variants of NSGA-II. Their findings validated the efficacy of these methods in minimizing total production costs, production time, greenhouse gas emissions, and hazardous waste generation [9]. Doerr B et al. proposed an ambulance fleet management approach grounded in the MILP model to mitigate the inadequacies in rescue vehicle capacity. Utilizing NSGA-II alongside an enhanced particle swarm optimization algorithm, they optimized the route sequence, thereby minimizing service completion time and patient deterioration rates. The results underscored the model's efficacy in ensuring timely medical assistance [10]. Babaeinesami et al. adopted NSGA-II for designing closed-loop supply chain networks, comparing its performance with the ϵ-constraint method. By fine-tuning parameters via the Taguchi design method, they improved NSGA-II's performance, enhancing solution time and generating effective Pareto solutions for efficient closed-loop supply chain network design [11]. Teo T T et al. proposed a fuzzy logic-based FEMS system for energy management in grid-connected microgrids featuring renewable energy and energy storage systems. NSGA-II was leveraged to optimize fuzzy membership functions, enabling the selection and implementation of the optimal compromise solution in the controller. The findings demonstrated the superior performance of the proposed FEMS system at the simulation level compared to alternative control strategies [12].

In summary, a considerable body of research has been dedicated to exploring the applications of both Apriori and NSGA-II, yielding promising outcomes. Building upon this foundation, this study introduces an enhanced version of the Apriori algorithm. Additionally, NSGA-II is employed to address a dual-objective optimization risk control model, aiming to enhance the effectiveness of DSR evaluation and control. Through these enhancements, the study endeavors to achieve superior results in managing data security risks.

3 DATA SECURITY RISK ASSESSMENT AND CONTROL BASED ON IMPROVED APRIORI AND NSGA-II

For the evaluation of network DSR, this study introduces an enhanced version of the Apriori algorithm and delineates its process comprehensively. Leveraging the Apriori methodology, a risk function tailored to the data structure of DSR evaluation is formulated. Additionally, NSGA-II is adopted to tackle the multi-objective optimization problem inherent in risk control. Subsequently, a maximum satisfaction approach is employed to derive the optimal compromise solution.

3.1 Data security risk assessment based on improved Apriori

In DSR evaluation, mining Association Rules (AR) within the dataset enables the identification of correlations between different data items and the detection of abnormal patterns or behaviors. By scrutinizing and analyzing these abnormal patterns, potential security threats can be promptly identified and addressed [13, 14]. AR entails uncovering specific patterns or relationships among multiple variables in data, which often exist implicitly in daily life and production. While individuals can intuitively perceive correlations in small datasets through experiential thinking, identifying AR in large datasets necessitates adherence to specific rules. Figure 1 depicts the fundamental model of AR.

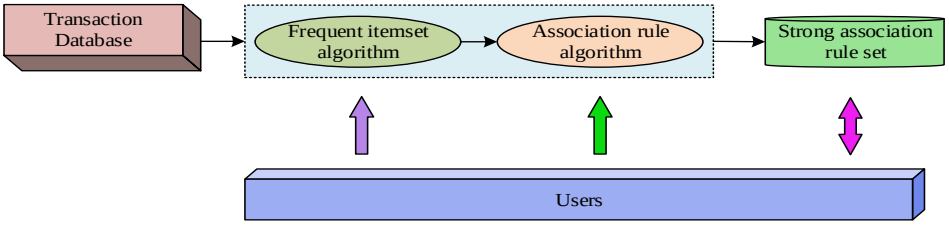

Fig.1 Basic model of the association rules

Apriori , a seminal data mining algorithm introduced by Rakesh Agrawal and Ramakrishnan Srikanth, primarily employs a breadth-first search strategy for exploration. It systematically excludes branches that fail to meet predetermined conditions by establishing prior criteria, ultimately yielding desired results [15]. Apriori is characterized by three main attributes. Firstly, it efficiently extracts data pertinent to the target objective and condenses three-dimensional data, encompassing factors, transaction numbers, and outcomes, into two-dimensional data, retaining only the former two dimensions. Secondly, Apriori predominantly operates on the dimensionality-reduced data, effectively mitigating computational burdens. Finally, during calculations, it selectively prunes data columns while maintaining transaction quantities constant, thus rendering the calculated result interpretable as support. Assuming transaction Ti within transaction set TS is of order m*n, the resulting Association Rule (AR) is represented by equation 1.

$$Y \Rightarrow Z(S\%, C\%) \tag{1}$$

Apriori's support level is represented by equation (2).

$$Sup(Y \Rightarrow Z) = S\% = \sum_{i=1}^{m} \{(Y \Rightarrow Z) \subseteq T_i\} \tag{2}$$

The confidence level of Apriori is represented by equation (3).

$$Con(Y \Rightarrow Z) = C\% = \frac{\sum_{i=1}^{m} \{(Y \Rightarrow Z) \subseteq T_i\}}{\sum_{i=1}^{m} \{Y \subseteq T_i\}} \cdot 100\% \tag{3}$$

The Apriori algorithm proceeds through specific steps. Initially, it generates all possible individual itemsets from a given database and computes the support for each itemset. Subsequently, itemsets with support below the predefined threshold are discarded, yielding frequent itemsets with support equal to or greater than the threshold. Following this, elements within the frequent itemsets are paired to generate a candidate set. The support of each itemset within the candidate set is then calculated, and itemsets with support below the threshold are eliminated

to obtain the second frequent itemset. Next, the candidate items within the second frequent itemset are decomposed into individual elements, and any three-element combinations are merged to exclude superset candidate items that were previously discarded, resulting in a refined candidate set. Subsequently, the support of each itemset within the candidate set is recalculated, and itemsets with support below the threshold are pruned to derive the third frequent itemset. These steps are iteratively repeated. After a finite number of iterations, if the resulting frequent itemset is empty, the previous frequent item set represents the maximal frequent itemset obtained. Figure 2 illustrates the detailed process of the Apriori algorithm.

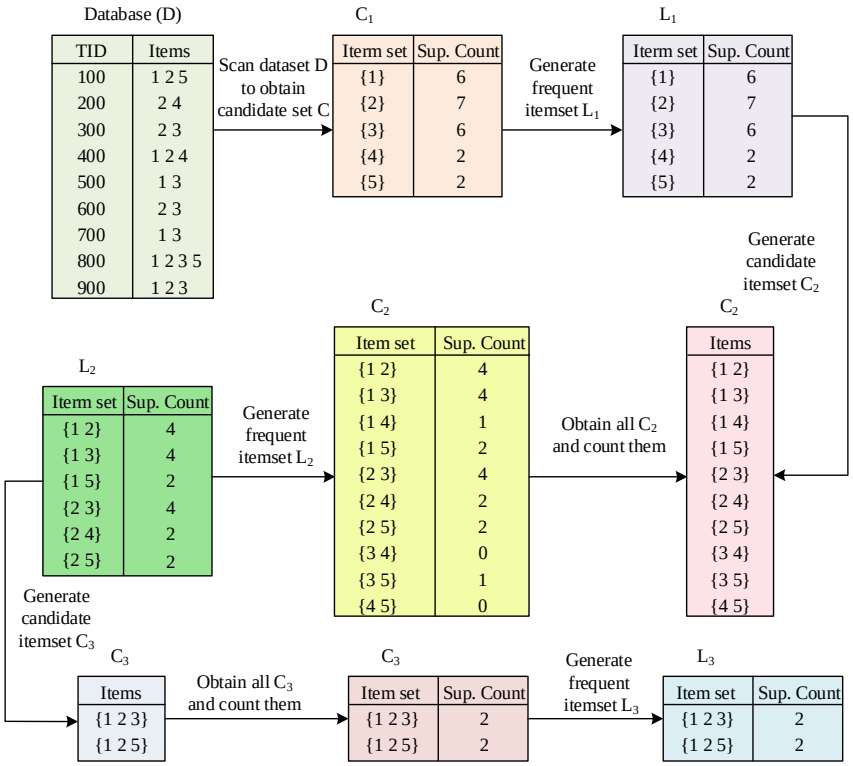

Fig.2 Specific process of Apriori

Apriori proves instrumental in capturing data correlations, thereby facilitating subsequent DSR evaluation and control. However, traditional Apriori suffers from drawbacks such as frequent database scans, generation of redundant intermediate terms, and limited sensitivity to initial data [16]. To address these limitations, an enhanced Apriori algorithm leveraging basic data partitioning and binary tree concatenation is proposed. The study primarily optimizes the original algorithm from three key perspectives. Firstly, a Boolean matrix representation is adopted for efficient computer recognition and operation, concurrently alleviating storage burdens on the database. Secondly, initial data are filtered to minimize the inclusion of irrelevant data, thus enhancing algorithmic efficiency. Furthermore, employing a block calculation approach, the Boolean matrix is partitioned into column terms, with frequent itemsets of each submatrix computed separately. Lastly, adhering to concatenation principles, the corresponding frequent itemsets are amalgamated into the maximal frequent itemset. In instances where the factor itemset is substantial, a binary tree approach is employed for computation, optimizing hierarchical architecture and achieving expedited convergence. Figure 3 illustrates the improved algorithm.

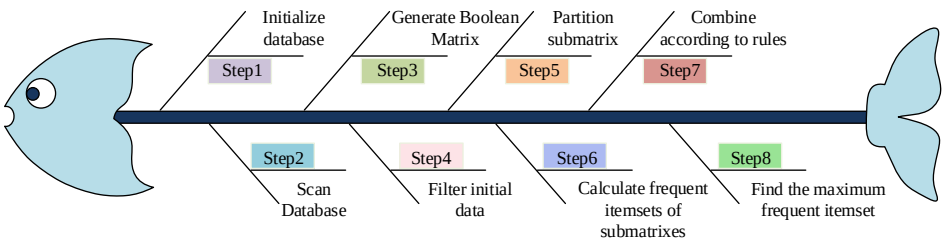

Fig.3 Flow chart for improving the Apriori

3.2 Data security risk control based on NSGA-II

The primary objective of DSR control is to maintain the risk level of network data at a secure threshold while implementing varying degrees of control measures for different risk sources to minimize associated costs. Utilizing Apriori, a risk function is formulated tailored to the data structure of DSR evaluation. In addition to ensuring effective management of network DSR levels, the cost of risk mitigation is integrated into the control objectives, with analyses conducted to ascertain cost disparities across different risk sources. Subsequently, a dual-objective control model is developed to simultaneously minimize risk levels and mitigate cost consumption [17]. The objective function aiming to minimize the DSR level is represented by equation 4.

$$\min P = \max_{k=1}^{6}(P^k) \tag{4}$$

In equation 4, k represents the total quantity of risk types. P means the overall risk state value of network data. P^k means the k risk state value. The objective function with the lowest cost consumption of risk k is represented by equation 5.

$$\min C^k = C_m^k \sum_{i}^{n} c_i (1 - \frac{x_i}{\alpha(p_i)}) \tag{5}$$

In equation 5, n represents the total number of risk sources corresponding to risk k. C^k is the cost of controlling risk k. C_m^k means the cost of controlling all risk sources for risk k. c_i refers to the cost of controlling the i-th risk source relative to other risk sources. $\alpha(p_i)$ represents the state value before the control of the i-th risk source. Risk k's calculation function is represented by equation 6.

$$P^k = f(x_i) \tag{6}$$

The constraint conditions are represented by equation 7.

$$\begin{cases} 0 \le P^k \le P_m^k \\ 0 \le C^k \le C_m^k \\ 0 \le x_i \le \alpha(p_i) \\ 1 \le i \le n \\ k = 1, 2, ..., n \end{cases} \tag{7}$$

When controlling network data risks, certain conditions must be met. Firstly, the post-control risk status value should not exceed the pre-control risk status value. Secondly, the total cost incurred must not surpass the total cost of controlling all risk sources. Additionally, the post-control state value of a risk source should not exceed its pre-control state value. Furthermore, the number of controlled risk sources should not exceed the total number of risk sources corresponding to that risk type. Leveraging the characteristics of the dual-objective risk control model, NSGA-II is introduced to address the dual-objective optimization risk control model. NSGA-II builds upon NSGA by incorporating the concepts of fast non-dominated sorting, congestion degree, congestion comparison operators, and elite strategies [18]. Congestion degree denotes the density of individuals surrounding a point within a given population, denoted as id. id is represented as the minimum rectangle that solely encompasses individual i without including other individuals, thereby circumventing errors stemming from decision makers specifying shared radii in NSGA [19]. The specific calculation steps for id are outlined as follows: Firstly, the distances between individuals within the same layer are initialized, as depicted in equation 8.

$$H[i]_d = 0 \tag{8}$$

Next, individuals in the same layer are arranged in ascending order according to the m-th objective function value, represented by equation 9.

$$H = sort(H, m) \tag{9}$$

Then, a large number M is given to avoid individuals on the edge being abandoned, and the specific calculation

is represented by equation 10.

$$H[0]_d = H[l]_d = M \tag{10}$$

In equation 10, l represents the marginal individual. Subsequently, individuals' crowding distance in the middle is computed using equation 11.

$$H[i]_d = H[l]_d + (H[i+1]_m - H[i-1]_m) \tag{11}$$

In equation 11, $L[i]_m$ stand for the i-th individual's m-th objective function value. The multi-objective optimizing problem's crowding distance is represented by equation 12.

$$H[i]_d = \frac{(H[i+1]\cdot f_1 - H[i-1]\cdot f_1)}{f_{1\max} - f_{1\min}} + \frac{(H[i+1]\cdot f_2 - H[i-1]\cdot f_2)}{f_{2\max} - f_{2\min}} \tag{12}$$

In equation 12, f represents the objective function. $f_{\max}$ and $f_{\min}$ are the objective function's maximum and minimum values, respectively. For optimization problems with q objectives, equation 13 is individual i's crowding distance.

$$H[i]_d = \sum_{r=1}^{q} \frac{(H[i+1]\cdot f_r - H[i-1]\cdot f_r)}{f_{r\max} - f_{r\min}} \tag{13}$$

Finally, the above operations are repeated for different objective functions. Figure 4 is a schematic diagram of population crowding degree.

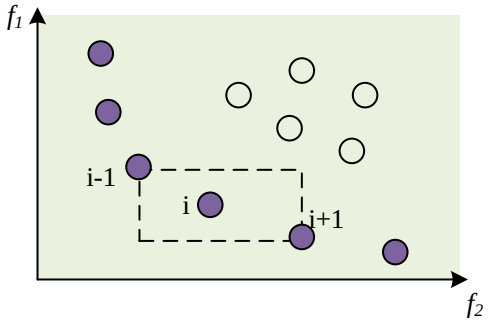

Fig.4 Schematic representation of the population crowding degree

Figure 4 illustrates the relationship between the crowding degree near individual i and id. A large id indicates a relatively crowded population. Introducing a crowding comparison operator helps maintain diversity and uniform distribution across a broad spectrum, thereby ensuring NSGA-II exhibits strong convergence performance on the Pareto surface. During individual selection, if two individuals have different id values, the one with the smaller sorting number is chosen. Conversely, if two individuals possess the same id value, the one with a less crowded surrounding environment is selected. Upon meeting these conditions, the algorithm executes, and the iterative calculation results are generated. Figure 5 depicts the process of NSGA-II.

Following the acquisition of the Pareto optimal solution set, a maximum satisfaction method is employed to derive the optimal compromise solution [20]. Each objective function's satisfaction, corresponding to each Pareto solution, is calculated utilizing the membership function. The standardized satisfaction of each non-dominated solution is represented by equation 14.

$$\mu^a = \frac{\sum_{b=1}^{B} \mu_b^a}{\sum_{a=1}^{A}\sum_{b=1}^{B} \mu_b^t} \tag{14}$$

In equation 14, μ^a represents the a-th non-dominated solution's standardized satisfaction. A refers to the quantity of non-inferior solutions. B means the quantity of objective functions.

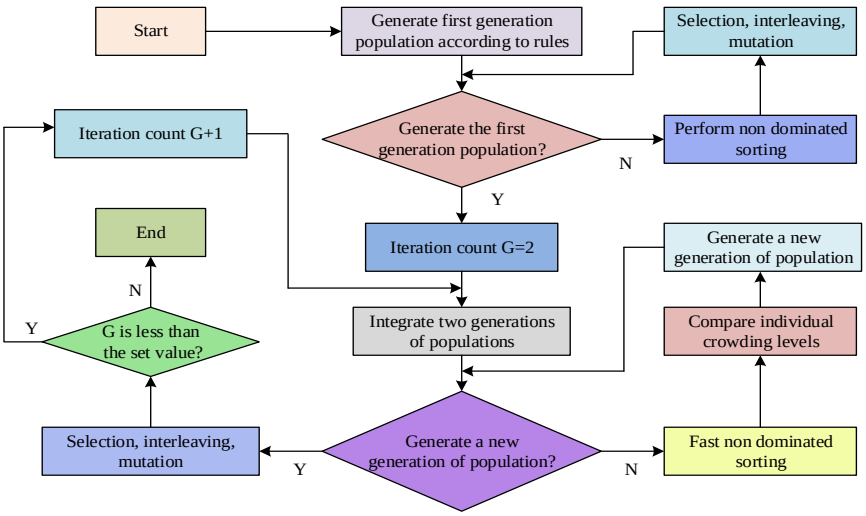

Fig.5 Flowchart of NSGA-II

The non-dominated solution with the highest standardized satisfaction value is considered a compromise solution. The dual objective functions used in this study are risk level and cost consumption minimization, therefore a slightly smaller fuzzy satisfaction function should be chosen, so that the objective function value is smaller, the fuzzy satisfaction function value is closer to 1. The small-scale fuzzy satisfaction function is represented by equation 15.

$$\mu_b = \begin{cases} 0, & f_b \geq f_{b\max} \\ 1, & f_b \leq f_{b\min} \\ \dfrac{f_{b\max} - f_b}{f_{b\max} - f_{b\min}}, & f_{b\min} \leq f_b \leq f_{b\max} \end{cases} \tag{15}$$

In equation 15, f_b refers to the function value for solving the b-th objective function. $f_{b\max}$ and $f_{b\min}$ stands for this function's maximum and minimum values in the solution set, respectively. Figure 6 shows a slightly smaller satisfaction function curve.

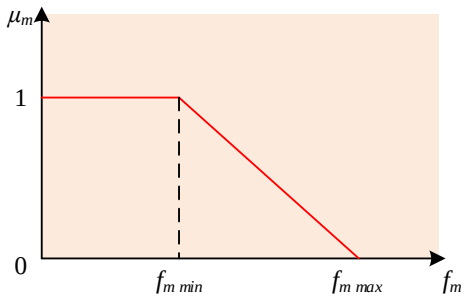

Fig.6 Parsmall satisfaction function curve

4 EXPERIMENT RESULTS

The study commenced by analyzing the efficacy of enhancing Apriori, with particular emphasis on the algorithm's spatial consumption and computational efficiency. Practical application outcomes were then validated. Subsequently, the Pareto frontier distribution results of NSGA-II were evaluated using the DTLZ1 and DTLZ2 testing functions. The efficacy of the dual-objective risk control function was verified, ultimately leading to the determination of the optimal risk control scheme.

4.1 Data security risk assessment analysis based on improved Apriori

To verify the effectiveness of the enhanced Apriori algorithm, a comparative analysis was conducted using MATLAB. Each item was assumed to occupy 1 storage space in the intermediate item set. By manipulating variables such as the number of transactions (T), the number of element items (Co), and the minimum support (minSup), the spatial consumption of Apriori before and after enhancement was compared. Figure 7 illustrates the comparison of temporary storage space required by Apriori before and after improvement.

In Figure 7(a), it is observed that as the number of transactions increased, the maximum number of intermediate itemsets in Apriori also increased. When T=110, the maximum value could reach up to 5000. Conversely, the intermediate itemsets for the improved Apriori remained stable around 100. In Figure 7(b), as the number of elements increased, the maximum number of intermediate itemsets in Apriori also increased. However, the intermediate itemsets in the improved Apriori remained stable around 0. In Figure 7(c), when the minimum support was set to 30, the maximum number of intermediate itemsets in Apriori decreased. Remarkably, the improved Apriori reduced the maximum number of intermediate itemsets to 0 when the minimum support was 24. Overall, the enhanced Apriori algorithm demonstrated reduced space consumption and improved efficiency compared to the traditional Apriori approach.

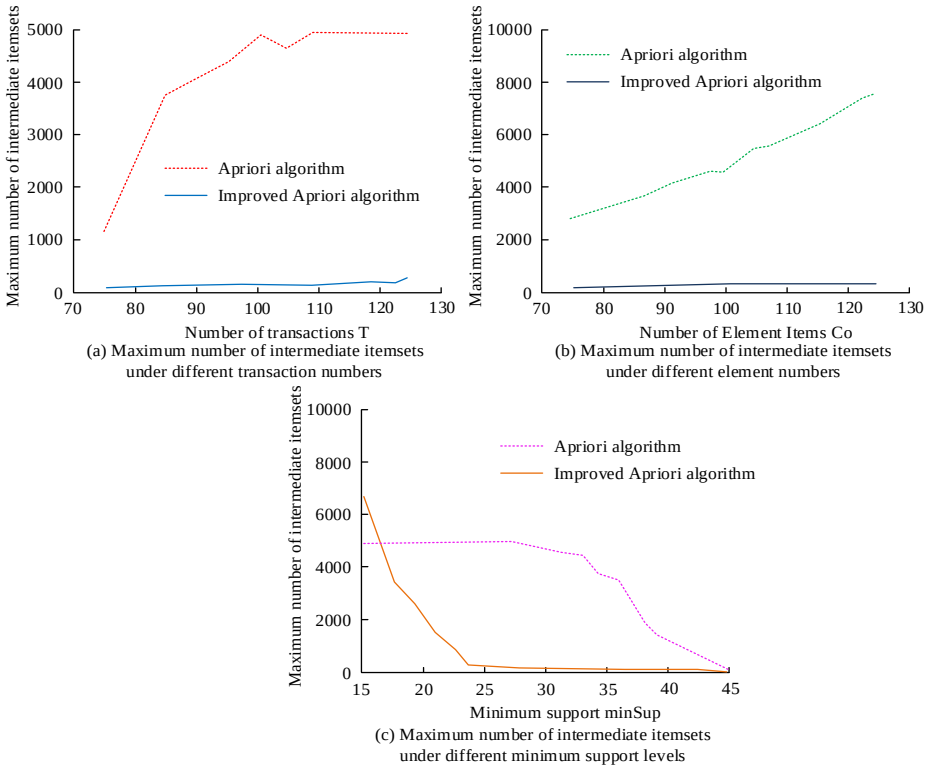

Fig.7 Comparison of temporary storage space required for Apriori before and after improvement

Continuing in a consistent environment, the study utilized the Northwind instance database as the data source and employed the C# programming language to conduct empirical analysis and comparative testing on both Apriori and the enhanced Apriori algorithm. Figure 8 illustrates the comparison of running times between the two algorithms.

As depicted in the figure, the running time of each algorithm increased with the number of data records. Notably, the running time of the improved Apriori was consistently lower than that of Apriori, and its growth rate of running time was also lower. For instance, when the data records reached 500, the running time of the improved Apriori was approximately 6×10^4 ms, whereas the running time of Apriori surged to as high as 8×10^4 ms. These findings indicate that the improved Apriori algorithm exhibits higher computational efficiency compared to the traditional Apriori approach.

The study further utilized 1800 sets of safety evaluation data from a specific railway unit in China spanning from 2021 to 2022 for risk analysis. Following data preprocessing, a foundational dataset was constructed. The minimum support was set to 20% of the total data, i.e., min Sup=360, while the minimum confidence was set to min Con=50%. Table 1 presents the top 10 risk factors and their corresponding confidence levels for network security within the railway unit.

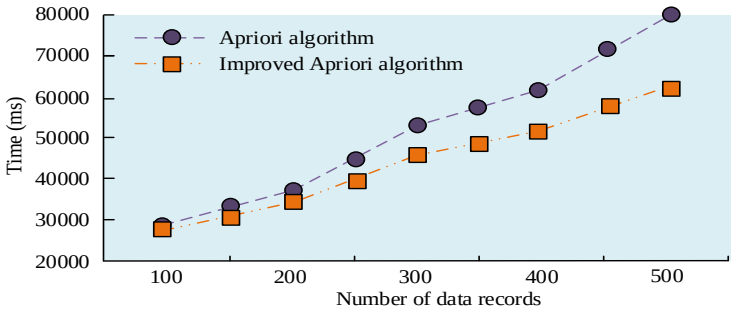

Fig.8 The running time of Apriori before and after improvement

The analysis revealed that the probability of the main risk factors affecting network security within this unit exceeded 0.7, with the confidence level of association between various factors and medium to high risks hovering around 75%. These calculation results were largely consistent with the ”main network security risks” outlined in the unit’s 2021 network security work summary. This alignment serves to verify the accuracy and reliability of the calculation outcomes.

Table.1 Risk factors and confidence of the top 10 network security ranking of a railway unit

Indicator items	Support level	Probability of occurrence	Medium to high risk confidence
Data remote disaster recovery	1451	0.7899	0.7507
Password product procurement	1412	0.7851	0.7631
Development management	1405	0.7612	0.7432
Network security plan review	1317	0.7513	0.7436
Regulatory system integrity	1304	0.7364	0.7512
Host device identity authentication	1294	0.7269	0.7491
Emergency resource configuration	1289	0.7133	0.7439
Emergency plan	1284	0.7042	0.7691
Application system identity authentication	1281	0.7012	0.7439

4.2 Analysis of data security risk control based on NSGA-II

To assess the effectiveness of NSGA-II, the studyconducted a performance test and applied it to the DTLZ1 and DTLZ2 functions, obtaining their Pareto frontier results as depicted in Figure 9. In Figure 9(a), NSGA-II demonstrated a relatively uniformand convergent Pareto front in theDTLZ1 test function. This suggests robustness and convergence of the algorithm. Similarly, in Figure 9(b), NSGA-II’s Pareto front distribution in DTLZ2 appeared relatively even, further indicating its stability and convergence performance. Overall, these findings highlight the efficacy of NSGA-II in generating well distributed and convergent Pareto fronts formulti-objective optimization problems.

The study proceeded to compare the performance of the Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) with NSGA-II. To facilitate a comprehensive comparison of the dual objective risk

control Pareto frontier, the risk level and cost consumption function values were normalized using the application system and data security of a railway unit as an example.

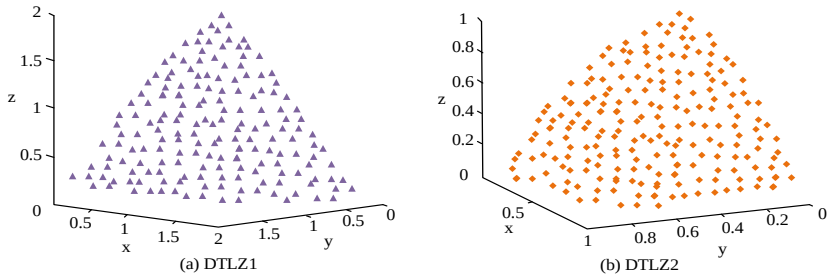

Fig.9 Pareto frontiers for different test functions

Figure 10 illustrates the Pareto frontier results for dual objective riskcontrol solved by the two algorithms. In Figure 10(a), it is evident that the distribution of the Pareto optimal solution set solved by MOEA/D was relatively uneven, with the dual objective function values being higher compared to those obtained by NSGA-II. Conversely, in Figure 10(b), NSGA-II's Pareto optimal solution set exhibited favorable diversity and uniform distribution. This underscores its robust global search capability when addressing the constructed risk control model. NSGA-II effectively explores a broad solution space and yields a non-dominated solution set with significant diversity.

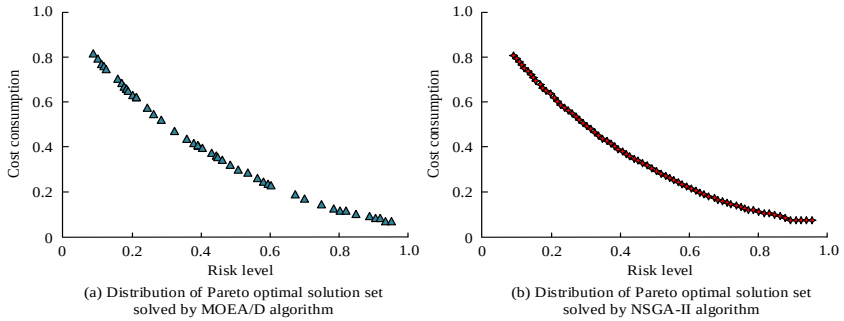

Fig.10 Pareto frontiers of dual objective risk control with different algorithms

Furthermore, from Figure 10, it can be observed that all non-dominated solutions form a curve that approximates the inverse function, where the values of these two objective functions are approximately inversely proportional. This trend indicates that optimizing one objective function inevitably leads to degradation in the other objective function.

The network DSR evaluation results were further analyzed, identifying risk analysis indicators relevant to the application system and data security of a specific railway unit. These indicators included E1 application system security configuration, E2 monitoring and defense, E3 application system identity authentication, and E4 important data security protection. Using MATLAB, a set of Pareto optimal solutions was obtained through experimentation. Subsequently, four Pareto frontiers with risk values close to 0.3, 0.4, 0.5, and 0.6 were selected, yielding four application systems and DSR control schemes.

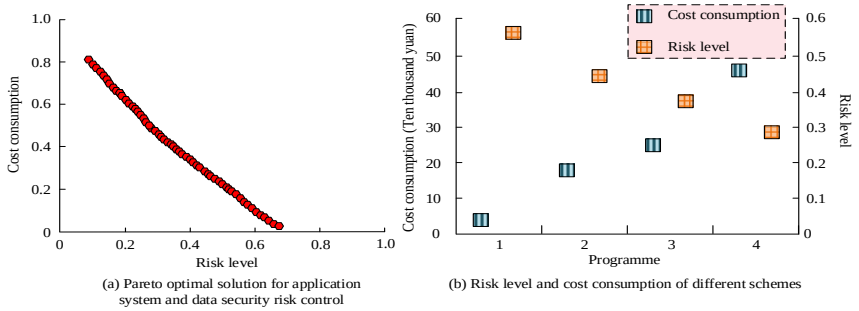

Fig.11 Pareto optimal solution with the four control scheme results

Figure 11 illustrates the Pareto optimal solution and the fourcontrol schemes. In Figure 11 (a), a clear inverse correlation between risk level and cost consumption is observed. Conversely, in Figure 11 (b), it is evident that as the risk level decreases, the cost consumption increases. Considering both risk state value and control cost, this paper adopts a third approach to control the application system and DSR. Inthisstrategy, the risk level is merely 0.386, with a corresponding cost consumption of only 2,404,800 yuan.

CONCLUSION

The study introduced improved Apriori and NSGA-II for coping with network DSR, yielding promising results. Specifically, as the transactions increased, traditional Apriori exhibited a corresponding increase in the maximum number of intermediate itemsets, peaking at 5000 when T=110. In contrast, the improved Apriori maintained stability with around 100 intermediate itemsets. Similarly, with an increase in elements, traditional Apriori experienced an uptick in intermediate itemsets, whereas the improved Apriori remained stable at around 0. Moreover, when the minimum support was set to 30, traditional Apriori exhibited a decrease in the maximum number of intermediate itemsets, whereas the improved Apriori reduced the maximum number to 0 with a minimum support of 24. Additionally, the running time of the improved Apriori was lower compared to traditional Apriori, with a growth rate also lower. For instance, with 500 data records, the running time of the improved Apriori was approximately 6 × 104 ms, while traditional Apriori soared to 8 × 104 ms.

NSGA-II demonstrated a relatively uniform and convergent Pareto front in both DTLZ1 and DTLZ2. Notably, in NSGA-II's Pareto optimal solution set, the values of the two objective functions exhibited an approximately inverse proportional relationship. As the risk level decreased, cost consumption increased accordingly.

In conclusion, the proposed improved Apriori showcased high computational efficiency and low space consumption, while NSGA-II exhibited excellent performance in risk control. However, for complex risk control problems, NSGA-II's convergence speed may be slow, and it may lack flexibility in handling the weights and correlations of different risk factors. Future research endeavors could focus on optimizing this algorithm to better address such challenges in risk control.

ACKNOWLEDGEMENT

This research was supported by the Guangxi Power Grid Technology Project under Grant 046100KK52222001.

REFRENCES

[1] Qoniah I, Priandika A T. Analisis Market Basket Untuk Menentukan Asossiasi Rule Dengan Algoritma Apriori (Studi Kasus: Tb. Menara). Jurnal Teknologi Dan Sistem Informasi, 2020, 1(2): 26-33.
[2] Abidin Z, Amartya A K, Nurdin A. Penerapan Algoritma Apriori Pada Penjualan Suku Cadang Kendaraan Roda Dua (Studi Kasus: Toko Prima Motor Sidomulyo). Jurnal Teknoinfo, 2022, 16(2): 225-232.
[3] Mirmozaffari M, Shadkam E, Khalili S M, Kabirifar, K, Yazdani R, Asgari Gashteroodkhani T. A novel artificial intelligent approach: Comparison of machine learning tools and algorithms based on optimization DEA Malmquist productivity index for eco-efficiency evaluation. International journal of energy sector management, 2021, 15(3): 523-550.
[4] Santoso M H. Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom. Brilliance: Research of Artificial Intelligence, 2021, 1(2): 54-66.
[5] Hendrastuty N, Setyawati S. PENERAPAN ALGORITMA APRIORI PADA APOTEK SHAQEENA UNTUK MEMPREDIKSI PENJUALAN BERBASIS ANDROID. Jurnal Teknologi dan Sistem Informasi, 2023, 4(3): 302-312.
[6] Abidin Z, Amartya A K, Nurdin A. Penerapan Algoritma Apriori Pada Penjualan Suku Cadang Kendaraan Roda Dua (Studi Kasus: Toko Prima Motor Sidomulyo). Jurnal Teknoinfo, 2022, 16(2): 225-232.
[7] Styawati S, Nurkholis A, Anjumi K N. Analisis Pola Transaksi Pelanggan Menggunakan Algoritme Apriori. J-SAKTI (Jurnal Sains Komputer dan Informatika), 2021, 5(2): 619-626.
[8] Zheng W, Liu Y, Doerr B. A first mathematical runtime analysis of the Non-Dominated Sorting Genetic Algorithm II (NSGA-II).Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(9): 10408-10416.

[9] Khettabi I, Benyoucef L, Amine Boutiche M. Sustainable multi-objective process planning in reconfigurable manufacturing environment: adapted new dynamic NSGA-II vs New NSGA-III. International Journal of Production Research, 2022, 60(20): 6329-6349.

[10] Doerr B, Qu Z. Runtime analysis for the NSGA-II: Provable speed-ups from crossover.Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(10): 12399-12407.

[11] Babaeinesami A, Tohidi H, Ghasemi P, Goodarzian F,Tirkolaee E B. A closed-loop supply chain configuration considering environmental impacts: a self-adaptive NSGA-II algorithm. Applied Intelligence, 2022, 52(12): 13478-13496.

[12] Teo T T, Logenthiran T, Woo W L, Abidi K, John T, Wade N S, Taylor P C. Optimization of fuzzy energy-management system for grid-connected microgrid using NSGA-II. IEEE transactions on cybernetics, 2020, 51(11): 5375-5386.

[13] Prasetya T, Yanti J E, Purnamasari A I, Dikananda A R, Nurdiawan O. Analisis Data Transaksi Terhadap Pola Pembelian Konsumen Menggunakan Metode Algoritma Apriori. INFORMATICS FOR EDUCATORS AND PROFESSIONAL: Journal of Informatics, 2022, 6(1): 43-52.

[14] Zhao Y, Zhang C, Zhang Y, Wang Z, Li J. A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis. Energy and Built Environment, 2020, 1(2): 149-164.

[15] Bashkari M S, Sami A, Rastegar M. Outage cause detection in power distribution systems based on data mining. IEEE Transactions on Industrial Informatics, 2020, 17(1): 640-649.

[16] Bashkari M S, Sami A, Rastegar M. Outage cause detection in power distribution systems based on data mining. IEEE Transactions on Industrial Informatics, 2020, 17(1): 640-649.

[17] Liu X, Sun J, Zheng L, Wang S, Liu Y,Wei T. Parallelization and optimization of NSGA-II on sunway TaihuLight system. IEEE Transactions on Parallel and Distributed Systems, 2020, 32(4): 975-987.

[18] Li H, Wang B, Yuan Y, Zhou M, Fan Y. Scoring and dynamic hierarchy-based NSGA-II for multiobjective workflow scheduling in the cloud. IEEE Transactions on Automation Science and Engineering, 2021, 19(2): 982-993.

[19] Jafari V, Rezvani M H. Joint optimization of energy consumption and time delay in IoT-fog-cloud computing environments using NSGA-II metaheuristic algorithm. Journal of Ambient Intelligence and Humanized Computing, 2023, 14(3): 1675-1698.

[20] Su C, Liu Y. Multi-objective imperfect preventive maintenance optimisation with NSGA-II. International Journal of Production Research, 2020, 58(13): 4033-4049.