





## Computer Assistance Analysis of Power Grid Relay Protection Based on Data Mining

Na Suo <sup>1</sup>  and Zheng Zhou <sup>2</sup> 

<sup>1</sup>Department of Electrical Engineering, Zhengzhou Railway Vocational and Technical University, Zhengzhou, Henan 450000, China, [Suona06@163.com](mailto:Suona06@163.com)

<sup>2</sup>Department of Electrical Engineering, Zhengzhou Railway Vocational and Technical University, Zhengzhou, Henan 450000, China, [232557644@qq.com](mailto:232557644@qq.com)

Corresponding author: Na Suo, [Suona06@163.com](mailto:Suona06@163.com)

**Abstract.** Relay protection and fault diagnosis are one of the important guarantees for the safe operation of power systems. With the widespread application of digital technology in the power system, the power system has produced data that is growing at an exponential rate. For this reason, this paper proposes a computer-aided analysis system for relay protection based on data mining. Apply data mining technology to the relay protection fault information processing system, and use data mining technology to extract many potentially important factors, facts, and correlations contained in a large amount of fault information. First, analyze the characteristics of data mining technology and power grid protection, and design the system structure of the power grid data intelligent analysis system. Secondly, the relevance analysis of the data of grid relay protection adopts a class-based frequent item set mining algorithm, and encapsulates the frequent item sets into classes, which is convenient for the calculation of the minimum support and the minimum confidence, weighted synthesis of exponential smoothing analysis model and regression analysis model to obtain prediction algorithm model. Finally, the historical data of a 110kv substation is analyzed and the proposed analysis framework and algorithm are verified.

**Keywords:** computer assistance; grid relay protection; data mining

**DOI:** <https://doi.org/10.14733/cadaps.2021.S4.61-71>

### 1 INTRODUCTION

Through the relay protection intelligent analysis system based on the platform of big data, the centralized remote intelligent diagnosis and fault analysis of the intelligent substation protection system can be realized. Under the conditions of centralized protection of professional technical resources and human resources, the protection system can be comprehensively improved. Operation level, fault analysis level, intelligent diagnosis level, improves the reliability and safety of smart grid operation [2]. How to make full and effective use of these valuable historical data, adopt appropriate data analysis techniques, and conduct targeted analysis and mining on it, obtaining valuable knowledge from it and providing decision support for the safe operation of the power grid has become a problem to be solved. Data mining technology is the process of

extracting potential and valuable knowledge models or rules from massive data, that is, exploring and analyzing large amounts of data according to predefined goals, revealing the hidden rules, and further integrating its modeled advanced and effective technical process [3]. With the continuous development of mining technology, it has been widely used in various fields [4]. The application of data mining technology to the intelligent analysis of the data of grid relay protection is one of its important application fields.

In recent years, data mining specifically for the safe and stable operation of power systems has developed rapidly. Among them, Lu et al. scholars have made outstanding contributions in this regard. The ATDIDT (currently renamed PEPITO) power system safety and stability-mining software developed by them has been put into commercial release [5]. Yue et al. proposed an expert system for fault diagnosis and recovery processing in substations. It integrates two knowledge expression modes: rule and process [6]. Gaber et al. proposed the fault diagnosis of substation based on artificial neural network, which used the inherent fault tolerance of neural network itself, but did not conduct special fault tolerance research, so the fault tolerance of the network is very limited [7]. Sekar et al, in the power system substation, which requires accurate and fast online fault identification, proposes a method combining cause-effect network and fuzzy rule library, and uses the good parallel processing capability of cause-effect network to retrieve faulty components [8]. Yang et al. use the basic regression analysis method to analyze the fault and establish the prediction model. When other factors are relatively stable, the prediction accuracy is relatively high [9]. The automation construction of the power system has been carried out relatively early and the overall level is relatively high. For a typical large-scale power system, with the increasing development of various computer monitoring equipment, geographic information systems and management information systems, the data in the power system database has exploded [10]. Regression analysis method is a mathematical method for studying the dependent relationship between variables and variables. Given multiple sets of independent variables and dependent variables, the regression equation is formed by studying the relationship between the respective variables and dependent variables. At the same time, the application of various automation systems and the continuous improvement of the level of information technology make the data in the power system have the characteristics of a wide range of data sources, a wide variety of data, and poor data quality.

This paper firstly summarizes the data mining technology from the concept of data mining, mining mode, and related technologies used in mining, and propose the architecture of the computer-aided analysis system for power grid data. Secondly, it analyzes the correlation analysis and time series prediction of the data-mining model for intelligent analysis of power grid data. It focuses on the concept of association analysis and typical frequent item set mining algorithms. On this basis, a cluster-based frequent item set mining algorithm suitable for the computer-aided analysis system of power grid data is proposed; the forecasting methods and technologies are analyzed, and a comprehensive time series-forecasting model suitable for intelligent analysis systems is proposed. Then, correlation and timing analysis using the collected test data 110kv substation computer-aided analysis of the proposed system is the result of the prediction. Finally, conclusions and prospects are given.

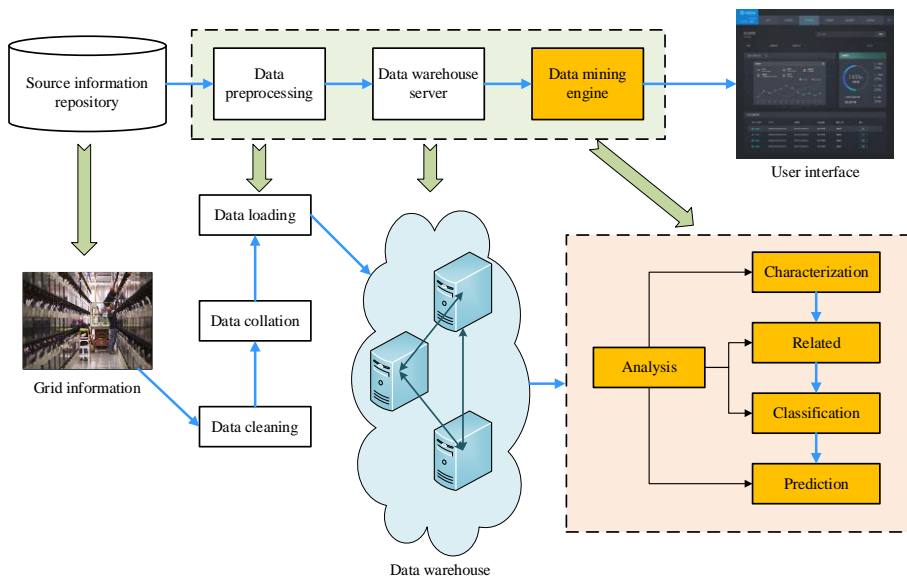
## **2 SYSTEM STRUCTURE AND TECHNICAL BACKGROUND OF POWER GRID RELAY PROTECTION DATA ANALYSIS SYSTEM**

### **2.1 Data Mining Technology**

Data mining is essentially a new information processing technology. Data mining technology improves people's application of data from online query operations to more advanced applications such as decision support, analysis and forecasting. This knowledgeable information can be used to guide advanced decision-making activities. From a narrow perspective, data mining can be defined as a process of extracting knowledge from a specific form of data set.

From a broad perspective, data mining is based on the fact that large data sets may be incomplete, noisy, uncertain, and in various storage forms. People do not know in advance of mining, hidden knowledge useful for decision-making process. Based on the broad view of data mining functions, the structure of a typical data mining system and its implementation steps are shown in Figure 1.

As shown in Figure 1, the source information repository is a database or a group of databases, data warehouses, spreadsheets or other types of information repositories for storing business information. Data cleaning, sorting and loading mainly preprocess information, the database or data warehouse server is responsible for extracting relevant data; the data mining engine is the core of the data mining part and consists of a set of functional modules for performing tasks such as characterization, association and correlation analysis, classification, prediction, and cluster analysis. The user interface communicates between the user and the data mining system to realize the interaction between the user and the system. In fact, data mining is a basic step in the knowledge discovery process. Knowledge discovery first extracts the data of interest from the data source and organizes it into a data organization form suitable for mining. Finally, the knowledge generated models to assess and integrate valuable knowledge to the intelligent system.



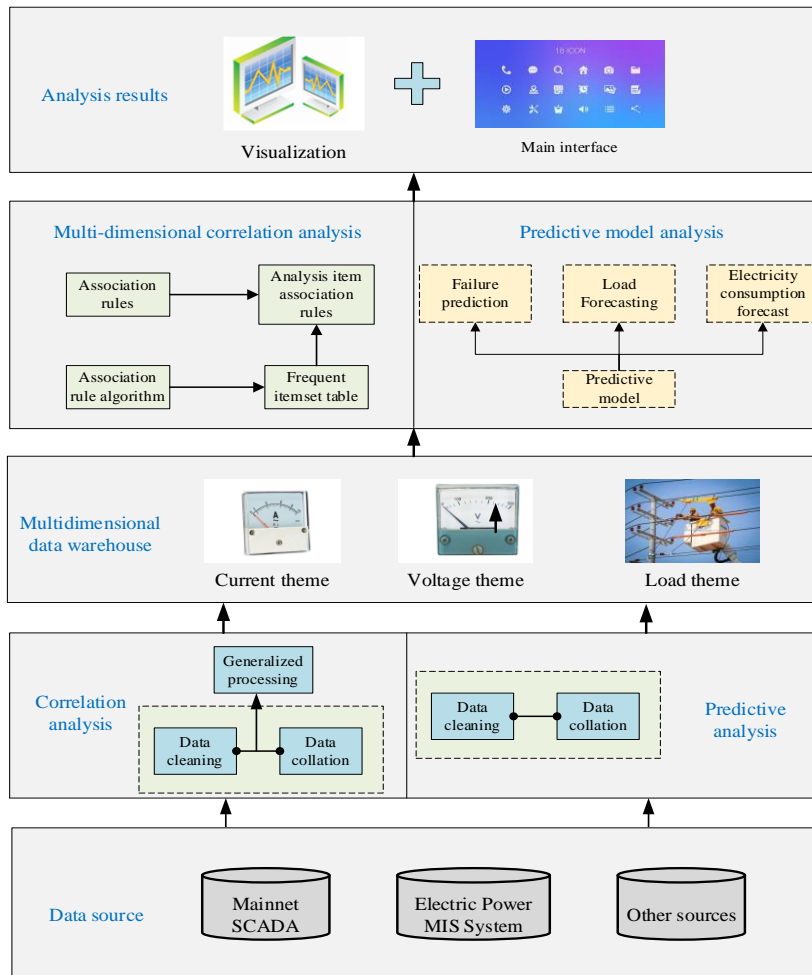
**Figure 1:** Data Mining System Structure.

## 2.2 System Structure of Data Analysis System for Power Grid Relay Protection

With the wide application of substation monitoring system, the function of dispatching automation system is different. With the increase of power interruption, the automation degree of power grid dispatching operation has reached a very high level. However, the automation of protection monitoring of relay, system failures and the analysis of relay protection action behavior, and relay protection management is relatively lagging behind, and its technical means can no meet the needs of the development of power grid dispatching operation management automation. In order to meet the requirements of grid dispatching, operation and management automation, the construction of relay protection, safety automatic devices and fault recorders based on modern network communication technology involves action devices when the power grid is abnormal. The fault information processing system, namely the "relay protection fault processing system", can not only realize the efficient use and full sharing of relay protection and related data and information, but also realize the analysis of fault information and the operation of relay protection. Making the relay protection professional Modernization of management is of great significance for improving the

normalization and intelligence of the dispatching system for safe operation of the power grid, and for improving the overall management of the dispatching operation of the power grid.

The computer-aided analysis of the system is mainly to conduct multi-dimensional correlation analysis and time sequence prediction on the historical data of grid relay protection. The reference value of historical data is generally valid in the range of 5 to 10 years. No matter how long it takes, it will lose its reference significance. Adopt multi-dimensional data warehouse storage based on relational storage. Its architecture is shown in Figure 2. In this system, a SQL Server 2016 database is used to store historical data accumulated by the power grid. The data-preprocessing module mainly completes the cleaning, sorting and loading of data in the business database.



**Figure 2:** Teaching parallel corpus construction design.

Since association, rule mining can discover the relationship rules between items or attributes and attributes that cannot be discovered by traditional methods, it has important research value. Multi-dimensional association rules are mainly embodied in discovering frequent patterns, correlations or causal relationships between classification attributes and decision-making attributes from massive data, to grasp the correlation characteristics between grid components from a macro perspective. Use the K-Means algorithm to discredited the data, divide the continuous attribute value into

discrete intervals, and divide the discrete attribute value into several different value ranges, thereby reducing the number of attribute values and improving the connotation of attribute values, it is convenient for the process of data mining and the visual display of mining results.

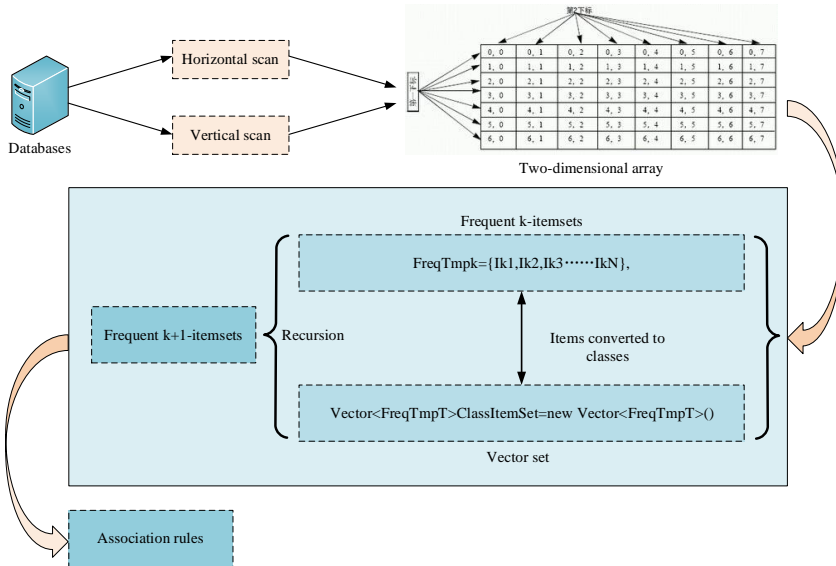
The fault diagnosis of relay protection is essentially a problem of pattern classification. It is necessary to establish fault phenomena and faults. Accurate mathematical models between the causes of obstacles are very difficult. This paper designs an intelligent analysis system to predict grid failures. First, extract the required data from the grid business data, and then use the K-Means clustering algorithm to discredited the data, establish a multi-dimensional data model data warehouse, and finally perform correlation analysis and Failure prediction

**3 ALGORITHM MODEL OF DATA ANALYSIS SYSTEM FOR POWER GRID RELAY PROTECTION**

**3.1 Data Protection Association Rule Mining Algorithm**

Association rule mining can discover the relationship rules between items and items or attributes and attributes that cannot be discovered by traditional methods. The problem of association rule mining is to specify the minimum support and minimum confidence in a given transaction number by the user. The process of finding suitable association rules in the database. Generally, the problem of association rule mining can be divided into two problems: one is to find frequent item set, and find all frequent item set through the minimum support min\_S given by the user, that is, item set with support not less than min\_S. The other is to generate association rules, and find association rules with no less than confidence in each maximum frequent item set given by the user.

Tian et al. improved an association rule-mining algorithm called Apriori based on previous work, which has been cited as a classic association rule-mining algorithm [11]. But it has two fatal performance bottlenecks: scanning the transaction database multiple times requires a large I/O load; it may produce a huge candidate set. In this paper, combining the characteristics of grid protection data, the classic Apriori algorithm is improved.



**Figure 3:** The processing process of frequent item set mining algorithm based on clusters.

Aiming at the shortcomings of the algorithm, the algorithm has been improved from the following aspects in order to realize the efficient mining of frequent item set of grid data. 1) Use classes to encapsulate frequent items, and the generated frequent item sets are stored in the form of classes. This class includes item names, support and confidence. Therefore, when calculating the support of candidate item sets, you can directly use the class items Obtained. 2) In order to overcome multiple scans of the transaction database, the performance bottleneck of a large I/O load is required, the data in the database is first stored in a two-dimensional array, and the frequent set mining operation is performed on the two-dimensional array, which avoids frequent transaction database scanning and improves mining efficiency.

Frequent set mining algorithm based on clusters. First, scan the transaction database and store the mining data in a two-dimensional array. Then conduct association mining on the data in the two-dimensional array, and the obtained frequent item set are stored in the class set in object format, and the calculation of support and confidence is realized by the class set. The processing process is shown in Figure 3.

### 3.2 Time Series Prediction Algorithm Mode

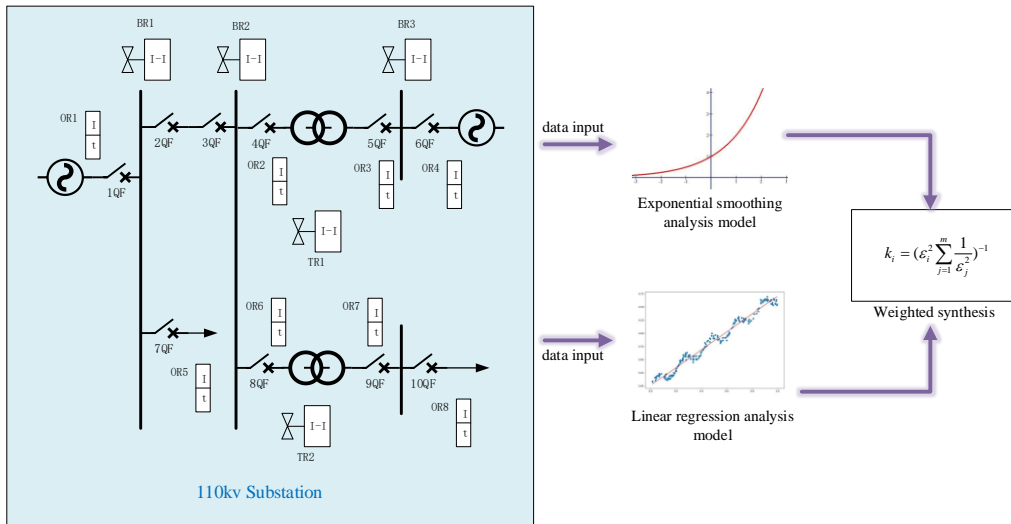
Forecasting is to make time-series forecasts on accumulated historical data, dig out future development trends, and provide a basis for macro decision-making. It is difficult for a single prediction model to guarantee satisfactory prediction results under any circumstances. Therefore, the power grid data intelligent analysis system adopts a comprehensive forecasting model, which combines the exponential smoothing analysis model and the linear regression analysis model to predict the data of power time series. By integrating different forecasting information and forecasting accuracy provided by different forecasting models, the forecasting results are optimized and the forecasting performance is improved. The comprehensive forecasting model is the appropriate weighted average of the different forecast values obtained by two or more different forecasting models. Finally, take its weighted average as a prediction model for the final prediction result. The predictive model used in the computer-aided analysis of the data of grid relay protection is shown in Figure 4.

In order to illustrate the effectiveness of the prediction algorithm model, a simple distribution network example is used to illustrate. As shown in the figure, there are four bus bars, two transformers, and three lines in the distribution network. Equipped with bus differential protection BR1 to BR4; current protection OR1 to OR8; transformer differential protection TR1, TR2. According to the operation principle of power grid protection, according to the various faults that occur during the operation of the power grid the protection and circuit breaker actions that may occur under the following conditions. Extract relevant data and perform correlation analysis, and input the obtained data into the prediction algorithm model to predict and analyze the fault information of grid relay protection.

The exponential smoothing analysis model and the regression analysis model simultaneously accept the original data, that is, the historical data sequence of the amount to be predicted to generate their own predicted values, and then these predicted values are weighted and integrated to obtain the final predicted data. The exponential smoothing analysis model is also called the exponential weighted average method. It is actually a weighted moving average method. A method selects the weight value of each option as the mean value of a decreasing exponential series. The exponential smoothing method solves the shortcomings of the moving average method that requires several observations and does not consider the data of the period before t-n. Through a certain averaging method, the random fluctuation in the historical statistical sequence is eliminated, and the main development trend is found. For time series  $x_1, x_2, x_3, \dots, x_n$ , the one-time smoothing index formula is:

$$F_t = \alpha x_t + (1 - \alpha)F_{t-1} \quad (1)$$

Where  $\alpha$  is the smoothing coefficient,  $0 < \alpha < 1$ ,  $x_t$  is the observed values of historical data sequence  $x$  at  $t$ , and  $F_t$  and  $F_{t-1}$  are the smoothed values at  $t$  and  $t-1$ .



**Figure 4:** Predictive model of analysis system.

The university regression model is suitable for prediction problems where the prediction object is mainly affected by a related variable and the relationship between the two is linear. The working procedure is as follows: The important step is to establish a university regression model. The university linear regression model is a mathematical equation  $\hat{Y} = a + bX$  used to analyze the linear relationship between an independent variable and a dependent variable.

A very important problem that needs to be solved in comprehensive forecasting is how to appropriately determine the weighted weights of each single forecasting method. Weights are also called coefficients, which determine the status or proportion of a single model in combined forecasting. This system uses linear combination forecasting to determine the weight of each individual forecast model.

With  $m$  single prediction methods, the error of the  $i$ -th prediction method is

$$\epsilon_i = \sqrt{\sum_t |F_{it} - F_{ot}|^2} \quad (2)$$

The idea of linear combination forecasting is to select the appropriate weight  $k_i$  to minimize the total error  $\epsilon = \sum_{i=1}^m k_i \epsilon_i$ . Among them,  $\sum_{i=1}^m k_i = 1, k_i \geq 0$  and  $i = 1, 2, \dots, m$ , combined with the least square method, can get the weight:

$$k_i = (\epsilon_i^2 \sum_{j=1}^m \frac{1}{\epsilon_j^2})^{-1} \quad (3)$$

After calculating the predicted value obtained by the exponential smoothing analysis model and the regression analysis model, the predicted value is obtained after weighting processing and comprehensive analysis.

## 4 ANALYSIS OF RESULTS

### 4.1 Correlation Analysis Test Results

The algorithm model is the core of data mining, and the corresponding algorithm model is selected for different mining tasks for knowledge mining. According to the characteristics of power grid data and the needs of knowledge mining, this computer-aided analysis system mainly considers the association and prediction of two algorithm models. Correlation analysis is mainly to find frequent patterns, correlations or causal relationships between classification attributes and decision attributes from the power grid relay protection data, to grasp the correlation characteristics between data elements from a macro perspective.

In order to verify the effectiveness of the proposed association algorithm for frequent item sets data mining based on clusters, the association analysis of a 110kv substation data is carried out. The six failure causes are numbered according to T1~T6, which respectively indicate: manufacturing quality, design defects, accidental contact by personnel, bad weather, improper maintenance, and others. Three types of faults O1~O3 are selected for correlation analysis, namely: transformer fault, protection device fault and protection secondary circuit fault. Use the association rules to analyze the provided fault data. When the minimum support min\_S is set to 0.2, the correlation analysis result of the fault cause and the fault type is shown in Figure 5.

It can be seen from Figure 5 that "manufacturing quality" and "protection device failure" are closely related, while "design defects" and "protection secondary circuit failure" are closely related. Through the analysis of the association rules, we can get the knowledge and information we need for the timely detection of faults and problem analysis.

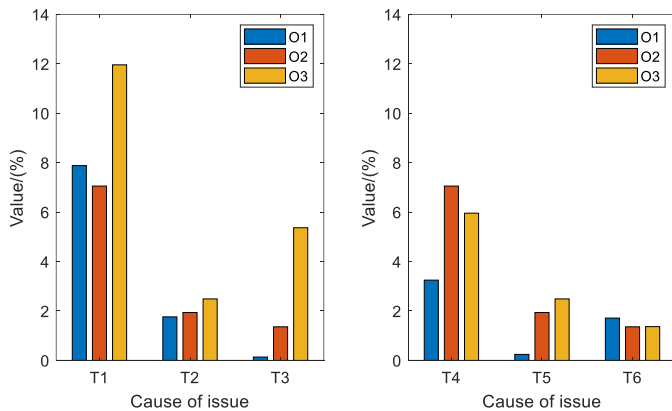
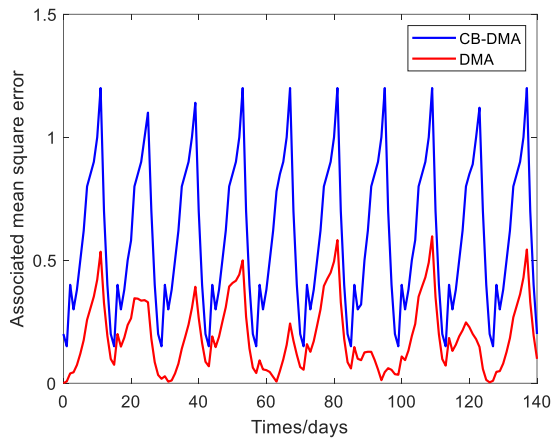


Figure 5: The correlation analysis result of the fault cause and the fault type.

In order to verify the superiority of the cluster-based frequent itemset data mining association algorithm (CB-DMA) proposed in this paper compared with the traditional non-cluster-based frequent itemset data mining association algorithm (DMA), the "manufacturing quality" and "protection device failure" are adopted. The correlation error is used to verify the algorithm. Using nearly half a year's failure data, the mean square error of the correlation matching between the failure cause and the failure type is analyzed. The analysis result is shown in Figure 6.

It can be seen from the figure that the classification of clusters in advance improves the matching accuracy of fault types and fault causes. The algorithm proposed in this paper has lower correlation mean square error.

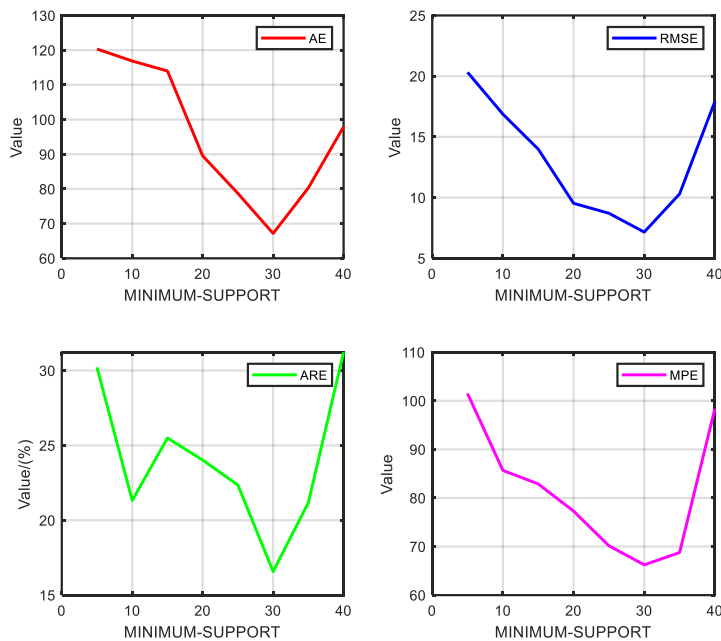




**Figure 6:** Mean square error results of different correlation algorithms.

**4.2 Time Series Prediction Test Results**

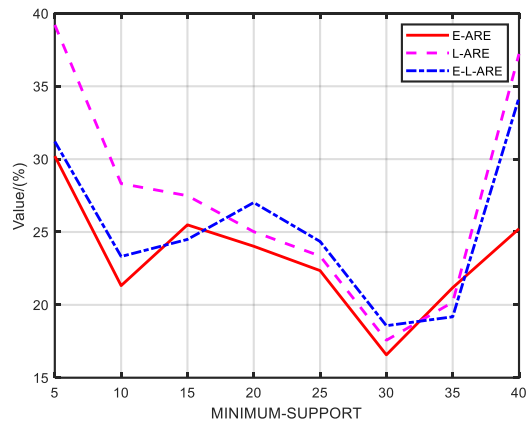
Here, only the daily data of transformer failure in an 110kv substation data is selected, the data of nearly one and a half months is selected to form the time series, and the data of the previous month is used as the model-training sample. Use nearly half a month's data to test the model's prediction effect and test the system. The main changed parameter is MINIMUM-SUPPORT, and the average error (AE), root mean square error (RMSE), average relative error (ARE), and maximum prediction error (MPE) are used as evaluation indicators. The prediction result is shown in Figure 7.



**Figure 7:** The prediction result of different error indicators.

According to the average relative error index, when the ARE range is between 10% and 20%, the system is good. The results show that when MINIMUM-SUPPORT is equal to 30, its prediction accuracy is a good fit.

In order to verify the superiority of the prediction model proposed in this paper, choose different main changes to set the parameter MINIMUM-SUPPORT to analyze its prediction average relative error (ARE) results. The average relative error results of the exponential smoothing prediction model, linear regression analysis model, and weighted comprehensive model are shown in Figure 8. It can be seen from the figure that the proposed prediction algorithm model is smoother than a single prediction algorithm, and has a smaller average relative error on the optimal MINIMUM-SUPPORT, which is more effective for the fault prediction of power grid relay protection.



**Figure 8:** ARE results of different prediction algorithm models.

## 5 CONCLUSION

After years of operation of the power grid, a large amount of sufficient fault data has been accumulated. How to use these data to find out the objective laws hidden after accidental failures is the focus of power workers' research. This paper discusses the framework model of computer-aided analysis of grid relay protection based on data mining, System function realization, key technology and testing in the research process. The proposed class-based frequent itemset mining algorithm has achieved good results in the analysis of the correlation of fault data. At the same time, the original regression linear prediction model is improved, and the exponential smooth prediction model and the regression linear prediction model are analyzed. It adopts the method of weighted synthesis to predict the possible causes of the failure. From the analysis results, the prediction accuracy of the proposed method is significantly improved. From the overall concept research and analysis of the power grid relay protection computer-aided analysis and development program and practice it with the data of a 110kv substation, it proves that this program is effective and provides a reference for the development of other analysis systems.

## 6 ACKNOWLEDGMENT

The 13th Five-Year Plan of Education Science of Henan Education Department in 2020 project name: Research on evaluation index of online teaching under the background of epidemic disease project number: 2020YB0547

Na Suo, <https://orcid.org/0000-0002-3561-6468>

Zheng Zhou, <https://orcid.org/0000-0003-2519-8386>

## REFERENCES

- [1] Amin, M.-S.; Chiam, Y.-K.; Varathan, K.-D.: Identification of significant features and data mining techniques in predicting heart disease, *Telematics and Informatics*, 2019, 36: 82-93. <https://doi.org/10.1016/j.tele.2018.11.007>
- [2] Chen, W.; Tsangaratos, P.; Ilija, I.: Groundwater spring potential mapping using population-based evolutionary algorithms and data mining methods, *Science of The Total Environment*, 2019, 684: 31-49. <https://doi.org/10.1016/j.scitotenv.2019.05.312>
- [3] Wang, L.; Li, Y.; Zhu, H.: Research on Transformer Relay Protection System Based on Ubiquitous Power Grid Technology[C]//IOP Conference Series: Earth and Environmental Science, IOP Publishing, 2020, 558(5): 052011. <https://doi.org/10.1088/1755-1315/558/5/052011>
- [4] Wu, D.; Li, P.; Zhou, H.: Research and Application of Intelligent Maintenance of Relay Protection Equipment Based on Internet of Things Technology, *E&ES*, 2020, 440(3): 032043. <https://doi.org/10.1088/1755-1315/440/3/032043>
- [5] Lu, J.; Zhang, L.: Data mining technology of computer testing system for intelligent machining, *Neural Computing and Applications*, 2020: 1-11. <https://doi.org/10.1007/s00521-020-05369-6>
- [6] Yue J.; Li C.: Type Identification and Classification of Cascading Failures of Time Sections of Dispatching and Operation Based on Data Mining, *MS&E*, 2020, 740(1): 012137. <https://doi.org/10.1088/1757-899X/740/1/012137>
- [7] Gaber, M.-M.; Aneiba, A.; Basurra, S.: Internet of Things and data mining: From applications to techniques and systems, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2019, 9(3): e1292. <https://doi.org/10.1002/widm.1292>
- [8] Sekar, K.; Mohanty, N.-K.: Data mining-based high impedance fault detection using mathematical morphology, *Computers & Electrical Engineering*, 2018, 69: 129-141. <https://doi.org/10.1016/j.compeleceng.2018.05.010>
- [9] Yang, X.; Wang, W.; Li, Y.: Analysis of Power Grid Fault Diagnosis Based on Association Rules and its Evaluation Index Optimization Method[C]//2018 Chinese Automation Congress (CAC), IEEE, 3092-3097. <https://doi.org/10.1109/CAC.2018.8623284>
- [10] Zhu, L.; Li, M.; Zhang, Z.: Privacy-preserving authentication and data aggregation for fog-based smart grid, *IEEE Communications Magazine*, 2019, 57(6): 80-85. <https://doi.org/10.1109/MCOM.2019.1700859>
- [11] Tian, M.; Zhang, L.; Guo, P.: Data Dependence Analysis for Defects Data of Relay Protection Devices Based on Apriori Algorithm, *IEEE Access*, 2020, 8: 120647-120653. <https://doi.org/10.1109/ACCESS.2020.3006345>