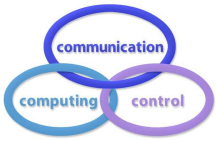
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 19, Issue: 3, Month: June, Year: 2024
Article Number: 6232, https://doi.org/10.15837/ijccc.2024.3.6232



CCC Publications

Enhancing Power Grid Data Analysis with Fusion Algorithms for Efficient Association Rule Mining in Large-Scale Datasets

Q. Q. Sun

Qiongqiong Sun
PingDingShan Vocational and Technical College,
Pingdingshan, Henan, China, 467000
Corresponding author: sunqiongqiong159@163.com

Abstract

Against the backdrop of the rapid development of information technology, the total amount of data has exploded, and efficient association rule mining methods for large-scale datasets have been studied. Conventional rule mining algorithms are subject to electrical constraints when working, and their convergence speed and data noise are currently the main problems they face. In order to accelerate the working process of the algorithm, this study introduces a data warehouse into the K-Means algorithm. The time series and voltage interaction functions are connected with the long-and-short-term memory network for efficient information analysis of power grid data, generating fusion algorithms. When mining association rules in electrical data sets, the pruning strategy is used to effectively reduce the search space, thus improving the efficiency of the algorithm. The pruning strategy applied in this study is confidence pruning, and the rules that do not meet the conditions are deleted by calculating confidence. For large-scale data sets, the distributed computing framework can dispatch tasks to multiple computing nodes, thus accelerating the computing speed. In the voltage interaction experiment, the storage and processing cost of electrical data is very high, so the research on using compression technology to reduce the dimension of storage space will reduce the computational complexity of the algorithm. The study conducts experiments on the Netloss dataset and three models, including long-and-short-term memory networks, to verify the superiority of the fusion algorithm. Under the same experimental voltage, the circuit power flows of the four models were 0.37, 0.64, 0.79, and 0.82A, respectively, indicating that the algorithm effectively controlled the electrical dataset. Its measurement accuracy was the highest among the four models, at 91.7%. The experimental results showed that the fusion algorithm proposed in the study had precise control ability in power grid datasets, and effectively mined association rules on large-scale datasets. This paper proposes a novel approach for efficient association rule mining in large-scale power grid datasets. A fusion algorithm is developed combining clustering, neural networks, and association rule learning. The technique incorporates data warehousing, time series modeling, and voltage interaction analysis to enhance information extraction from electrical data. Experiments on the Netloss dataset demonstrate the algorithm's effectiveness for power flow control and measurement accuracy compared to standard methods. Results show significant improvements in active power loss and network loss metrics as well. This research provides an important foundation for scalable analytics in smart power systems.

Keywords: Large-scale datasets, Efficient correlation, Rule mining, Data warehouse, Time series, Information analysis, Voltage interaction.

1 Introduction

The power system is a complex system that contains a large number of electrical equipment and sensors, which generate a huge amount of data [1, 2]. As a large-scale dataset, electrical datasets contain a large amount of valuable information that can have a significant impact on the stable operation of the power system. The commonly used data mining technique is association rule mining, which can discover the association relationships between data from large-scale datasets. With the increase of electrical data, the K-Means (KM) algorithm has the problem of low efficiency when processing large-scale electrical data [3]. In order to make it run quickly, the method proposed by experts is to establish a data warehouse in KM [4]. However, this method has a significant impact on the circuit voltage and faces problems such as low efficiency during operation. In order to enhance the global mastery ability of the algorithm, this study established a time series in the Long-and-Short-Term Memory Network (LSTM) for efficient information analysis of power grid data loss. Simultaneously considering voltage interaction in the circuit can greatly reduce the impact of current amplitude and generate a fusion algorithm (DWKM-LSTMIA). The main content of the study can be divided into four parts. The first part mainly analyzes and summarizes the application of the current KM. The second part introduces the connection method between Information Analysis (IA) and LSTM and introduces it to Data Warehouse combined with KM algorithm (DWKM). The third part conducts simulation experiments on the Netloss dataset. The last part analyzes and compares the performance of this model with traditional models, and points out the shortcomings that still exist in the research. The theoretical significance of this study lies in its ability to maintain electrical equipment and thus obtain efficient association rule mining for large-scale datasets. Intended to reduce equipment maintenance time for users and create more economic benefits.

2 Related works

In the study of rule mining in data, research is widely distributed internationally [5, 6, 7]. Meesala et al. used feature analysis of social media comments for rule mining on this dataset. They first generated a corpus of association rules and applied a multi-objective flower pollination algorithm to discover association rules for user opinions. The experimental results showed that their multi-objective cat swarm optimization algorithm outperformed other existing methods in terms of confidence and computational time [8]. Kota et al. proposed a processing method for sentiment analysis datasets by combining LSTM and attention methods. They used unsupervised learning algorithms and used word embeddings for natural language processing. Their methods included embedding layers and convolutional layers with maximum pools and used metrics such as accuracy and recall to analyze method performance. The experimental results indicated that their method helped to reduce complexity, thereby facilitating the processing of long-sequence input texts [9, 10]. Liu et al. proposed a KM-based image dataset extraction method considering the effects of soil moisture and collection time. They used density peaks to install hyperspectral imaging cameras on a tripod and collected images of wheat under different soil conditions. Dichotomy and KM were used to classify the wheat grayscale image dataset. The experimental results showed that conventional methods were influenced by collection conditions, and the error distribution was relatively dispersed. However, their extraction performance was less affected by collection conditions, and the error distribution was concentrated [11]. Power grid analysis is a commonly used data mining technology in association rule mining, which is used to discover the operation mode of power grid data. However, there are some limitations in the operation of conventional methods. Power grid data usually come from different data sources, including missing values and wrong data. These problems will affect the reliability of mining results; Power grid data has the characteristics of high dimension and large scale, including a large number of variables and records. This leads to an increase in computational complexity, and ordinary methods need more computational resources. There are some hidden patterns in power grid data, which are not easy to be discovered by traditional mining methods. Therefore, more advanced technology is needed to mine the hidden information.

With the increasing number of rules in the data, conventional methods face the problem of local

optima, and the DWKM-LSTM method has gradually entered the vision of many scholars. Antonello et al. proposed a data-driven approach to identify rare functional dependencies between components of different systems of complex technological infrastructure from large-scale databases. Their method was based on binary data representation and uses association rule mining algorithms to discover dependencies between different system components. They applied the proposed method to collect large-scale datasets in simulated synthetic databases, and experimental results showed that their method was effective [12]. Zhang et al. considered the increasing amount of information in the Internet of Things and used big data rule mining to improve the interaction of blockchain communication technology. They further optimized the efficiency of rule mining in data while ensuring the stability of data transmission. In order to address the impact of transmission delay between nodes, they were based on the reliability of weight transmission. The experimental results showed that their method had good efficient performance in terms of concurrent communication time and communication rule mining depth [13]. Ling et al. studied the combination of different feature positions in subway lines in order to conduct rule mining on the line dataset on subway foundations. They analyzed the differences between a track slab and a bridge pier, and the experimental results showed that their method performed well in large-scale datasets [14]. There is complex professional knowledge in the field of power grid, which is very important for mining power grid data. The existing methods lack a deep understanding of the power grid field, which leads to some difficulties in interpreting the analysis results. The analysis and mining results of power grid data are a series of patterns, and these results are difficult to understand. For professionals in the field of power grid, it takes extra effort to explain these results. In order to overcome these limitations, the combination of domain expertise and the proposed algorithm technology is studied to improve the accuracy of the analysis results.

Through the research of numerous experts and scholars, it has been found that the application research of KM and LSTM is very popular, but there is still little research on large-scale datasets. This study groundbreaking links the two and holds significant importance in dataset processing.

3 Research method

Efficient association algorithms can quickly find relevant items in the dataset and establish association relationships between data. These algorithms are commonly used in the field of data mining, as they can quickly discover patterns in data and make predictions. The KM algorithm and LSTM are classic rule mining algorithms that gradually increase the itemset through iteration to find the frequent itemset. Therefore, this study will apply the improved two algorithms to rule mining in large-scale electrical datasets to find the rules of the dataset.

3.1 Association Algorithm Combining Clustering and Neural Networks

With the modernization and improvement of intelligence of power systems, the scale and complexity of power grid data are increasing, including real-time data from sensors. Power grid data are diverse, and they have the characteristics of time series data or unstructured data. The growth of power grid data stimulates the demand for efficient data rule mining, and traditional analysis methods can not meet the requirements of efficient analysis of large-scale power grid data. Therefore, this research develops a fusion algorithm based on K- means. Among commonly used clustering algorithms, KM is used to divide the dataset into different clusters. The goal of this algorithm is to minimize the sum of squares between data points and cluster centers, with the advantages of high computational efficiency and good scalability. When working, KM first groups the data and uses each cluster as a training sample. Then, the data in the cluster is used as input features, and the cluster labels are used as outputs. In the process of cluster formation, the relationship between the data in the cluster and the original data is shown in equation (1).

$$Cl_i = \left[\sum_{i=1}^{k} \frac{p_i(P)}{o_i(O)} \right] /k \tag{1}$$

In equation (1), $o_i(O)$ represents certain data in the KM-generated cluster, $p_i(P)$ is their abnormal condition at neighboring points, and the total number of data points is represented by k [15]. The

KM algorithm performs poorly on datasets with non-spherical clusters, and the selection of initial clustering centers is sensitive [16]. To address this issue, this study combines the KM algorithm with the RNN association algorithm, using KM to divide the dataset into multiple clusters, and then using each cluster as a training sample for CNN. This method can reduce nonlinear data errors and improve the generalization ability of the model. The connection method is shown in equation (2).

$$LI_j = \left[\sum_{i=1}^{m} (Cl_i - x_i)^2\right]^{0.5} \tag{2}$$

In equation (2), x_i represents the clustering center in the initial data processed by the KM algorithm, and the total number of clusters is recorded as m. In order to accelerate the working process of the algorithm and expand its search domain, research has been conducted on adding Data Warehouse Management (DWM) management to clustering. It is a system used for integrating and managing large-scale data, with the ability to aggregate thematic time-consistent data, as shown in Figure 1 [17]. The main goal of a warehouse is to integrate data from different data sources to facilitate complex analysis by users. It contains data from operational systems and other data warehouses, where data can be cleaned and loaded to ensure data consistency.

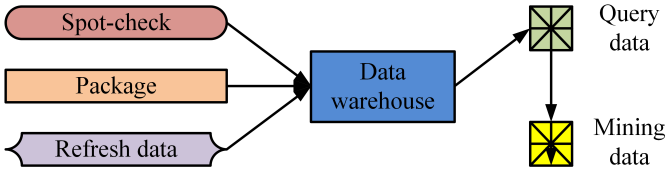


Figure 1: Diagram of data warehouse manager architecture

Figure 1 shows the internal structure of a data warehouse. From Figure 1, it can be seen that in the design of data warehouses, the scope and objectives of the data warehouse are first determined, and the logical model of DWM is manipulated [18]. Then data are extracted from various data sources and loaded into the data warehouse. The advantage of this method is to provide consistent data to support complex queries and analysis. In the process of data transfer, the logic between them follows the following equation (3).

$$Lo_t = \sum_{i=1}^{n} \sum_{y \in Lo} (Lo_i - LI_j)^2 \tag{3}$$

In equation (3) above, Lo represents the limiting condition of the data in that dimension, and Lo_i is an individual under that condition. The weighted sum of all individuals is denoted as Lo_t. In the time series of LSTM, various data are arranged in chronological order. Each data point is associated with a specific time point, which can be used to predict future trends and explore the correlation of data. This study applies the time series analysis of the LSTM part to mathematically model the data, as shown in equation (4).

$$\begin{cases} \alpha = \sigma * [\mathrm{Lo}_t{}^* h_t + a(i)b(x)] \\ \beta = \mathrm{tam} * [Zj * h_{t-1} + a(i-1)b(x-1)] \end{cases} \tag{4}$$

In equation (4), the data of the two input gates of LSTM is represented by $h_t, h_{t-1}, a(i), b(x)$ represents the model pattern running at the current time, and the link weight of the previous time node is recorded as $a(i-1), b(x-1). Lo_t, Z_j$ represent the parameters of two numbers, which are related to the environmental component [19]. By using this method, the improved KM can be connected to LSTM to form a composite model. In this model, the DWM method is introduced, as shown in Figure 2.

Figure 2 illustrates the method of connecting DWKM and LSTM. As shown in Figure 2, the fusion algorithm first preprocesses the time series data, including data cleaning and normalization operations, to ensure the quality and availability of the data. Then, an LSTM network is used to train the data to learn the temporal patterns and features of the data. LSTM can output its hidden layers as feature

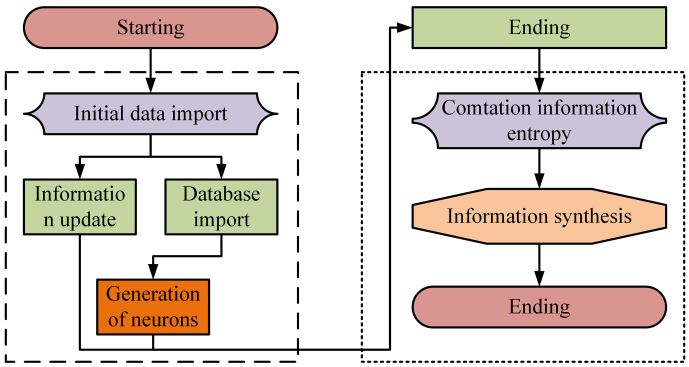


Figure 2: Working flow of the algorithm after connecting KM and LSTM with DWM method

representations based on historical data. In this process, each time step will correspond to the current index output, and its calculation method is shown in equation (5).

$$Ti(t) = Vy_t * \alpha + Vh_t * \beta + \gamma \tag{5}$$

In Equation (5), the weight generated by neurons during connection is recorded as Vy_t, Vh_t represents the propagation bias at the moment. The loss constant during the operation of the algorithm is expressed in terms of information loss between networks γ. With this method, the extracted features can be input into DWKM for clustering, and the cluster center can be calculated according to the weight of samples, and then the weight change of samples can be dynamically adjusted. According to the clustering results, the research makes full use of the modeling ability of the LSTM model for time series data and combines the dynamic weighted clustering ability of the DWKM algorithm to achieve clustering and analysis of time series data [20, 21]. The data modeling method is shown in Equation (6).

$$Mo(d) = \frac{\partial Ti(t)}{\partial L(t+1)} * \tan Co \tag{6}$$

In Equation (6), the loss amount of the output gate is expressed in $\partial L(t+1), Co$ represents the characteristic value of the new state of the data. Equation (6) can be used to establish the tree network diagram of the fusion algorithm DWKM-LSTM, which is a data structure for efficient mining of frequent patterns. It uses the tree to represent database schema and uses tree structure and path compression technology to improve mining efficiency. The advantage of this method is that the paths with the same prefix are merged, and the demand for storage space is reduced accordingly. The structural diagram is shown in Figure 3.

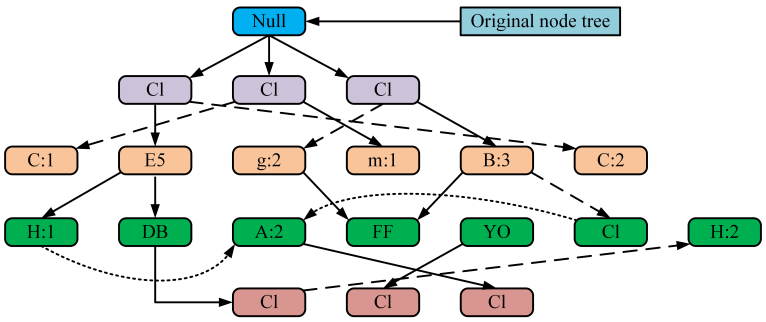


Figure 3: Workflow diagram of tree network system of DWKM-LSTM

Figure 3 illustrates the workflow of the tree DWKM-LSTM algorithm. As can be seen from Figure 3, this algorithm has four types of tree nodes, 23 in total. When the algorithm works, it first traverses transaction databases and sorts them in descending order of frequency. Then an empty root node is created and the items are inserted in the transaction into the algorithm tree. If the item already exists in a node in the tree, the count of the node is increased. Otherwise, it is placed in the corresponding

location. For each frequent item, its conditional mode base is studied and constructed, as shown in Equation (7).

$$Pb(c) = \frac{\partial Mo(d)}{\partial ootd} * \chi_t + \frac{\partial Hz(w)}{\partial (No)} \tag{7}$$

In equation (7), $Mo(d)$ represents the condition parameter in the mode base, ootd represents the position at this time, the flicker frequency of the algorithm is expressed in Hz(w), and the root node is recorded as No. Its value is highly sensitive and related to the calculation length χ_t. With this method, the subtree of each frequent item can be constructed recursively according to its conditional pattern base. The research starts from the leaf node and traverses the whole tree layer by layer. It can traverse the database once to build a mining mode and avoid the overhead of scanning the database many times. It is an efficient rule-mining algorithm and is widely used in power data [22]. The pseudo-code for the K-means algorithm involves calculating the distance between each sample and the cluster centers, selecting the nearest center as the cluster, updating each cluster's center to the average value of all samples in the cluster, and returning the sample clusters to LSTM for processing. The network then follows this pseudo-code. Additionally, the study focuses on the initialization of the weight matrix and bias vector in LSTM, examining the transition between the initialized hidden state and the memory state. Furthermore, it involves studying and calculating the forgetting gate and input gate for each time step in the input sequence and updating the memory state accordingly. Following the update of the hidden state, the sequence output for rule mining is obtained.

3.2 Efficient model construction of rule mining based on DWKM-LSTM algorithm

The goal of this algorithm is to improve the accuracy of the model by combining the advantages of clustering analysis and neural networks. In this regard, a neural network model is studied and constructed. The number of nodes in the input layer is equal to the dimension of the input feature, and the number of nodes in the output layer is equal to the number of categories of labels. Then, each cluster obtained by clustering is used as a training sample to train the neural network model. Firstly, the data set for training is prepared, and the neural network architecture is selected to build the model. It includes defining the activation function and loss function of the network and initializing the weight and bias of the model. Then the input data is propagated forward through the network and compared with the real tag. According to the value of the loss function, the parameters of the model can be updated. Then the steps of calculating loss and back propagation are repeated to gradually optimize the parameters of the model. Finally, the verification set is used to evaluate the performance of the model, and then the hyperparameters of the model are adjusted to further improve the performance of the model. For unknown data, the trained DWKM-LSTM model is used for prediction. The algorithm focuses on the selection and parameter setting of the clustering algorithm, as well as the adjustment of neural network structure and superparameters [23]. The power grid data is the power system load. These data are used to forecast the load and plan the power system, so as to help operators and managers understand the system operation. In order to extract the important information in detail, the DWKM-LSTM model is highly efficient through IA, as shown in Figure 4.

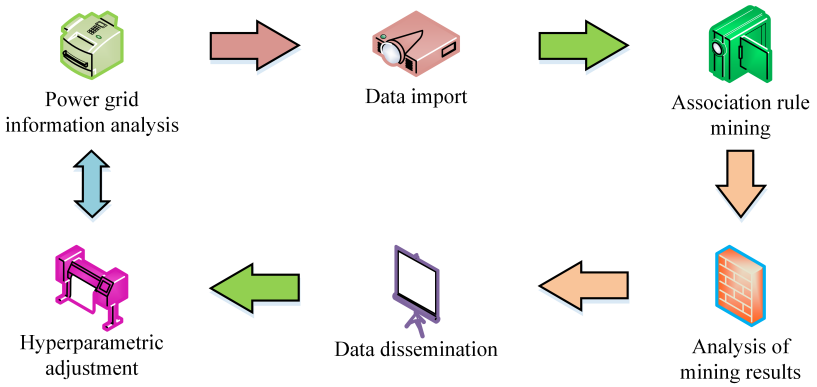


Figure 4: An efficient DWKM-LSTM model with information analysis method

Figure 4 shows the improved DWKM-LSTM efficient model process of IA. As can be seen from Figure 4, the research first preprocesses the electrical data set based on the fusion algorithm, and extracts features from the electrical data. Then, the extracted power grid data is converted into suitable rules, and the rules are converted into association rules at the same time. Then, according to the obtained association rules, the rule model of the electrical data set is constructed. Through the above steps, the algorithm can build an efficient electrical data set model [24]. In the calculation steps of power grid data, the calculation method used in this study is power flow calculation, which follows Equation (8).

$$\begin{cases} \varphi_1 = \eta_1 * \sum_{j=1}^{i} \left(Tr_1 - t_j^* \eta_j\right) + \sum_{k=1}^{K} \left(pV_1 + \kappa_j\right) \\ \varphi_2 = \eta_2^* \sum_{j=1}^{i} \left(Tr_1 - t_j^* \eta_j\right) + \sum_{k=1}^{K} \left(pV_2 + \kappa_j\right) \end{cases} \tag{8}$$

In equation (8), the power flow at two adjacent circuit points is recorded as φ_1 and φ_2, with η_1, η_2 representing their total power, and their active power and additional power are respectively expressed as η_j, κ_j. They are an important part of power system analysis, and their results can be used to evaluate the parameters at each node. pV_1, pV_2 are the circuit loads on two points, which are affected by the total power supply t_j. Assuming that t_j is known, the voltage amplitude in the circuit can be connected with the power, as shown in the following equation (9).

$$Va_1 = Va_0 + \eta_j^2 - \left(pV_1^2 + pV_2^2\right) \tag{9}$$

In equation (9) above, the circuit voltage amplitude under power on and power off conditions is recorded as Va_1, Va_0 respectively. This method considers the topology of the power system and is a simplified method for circuit power flow calculation. All elements in the power system are fixed value elements, and the environmental characteristics outside the power system are ignored [25]. Since there is distribution network loss in the circuit, equation (10) is studied and established to calculate the network loss of the circuit in order to cope with scenarios with high calculation speed requirements.

$$\begin{cases} Nl = t_j + \Delta Uk_t \\ \Delta Uk_l = \sum_{i=1}^{n} \Delta Sp / \Delta Ba \end{cases} \tag{10}$$

In Equation (10), ΔUk_l represents the total loss of the circuit at the current time, $\Delta Sp, \Delta Ba$ represent the gateway voltage and the balance voltage of the circuit. This calculation method is more accurate in calculating the circuit power flow. It considers the characteristics of the power system and can specifically face each node of the power system [26]. Distribution network loss calculation is the process of analyzing the power loss of the distribution network, and can effectively target the circuits with large losses. Its process is shown in Figure 5.

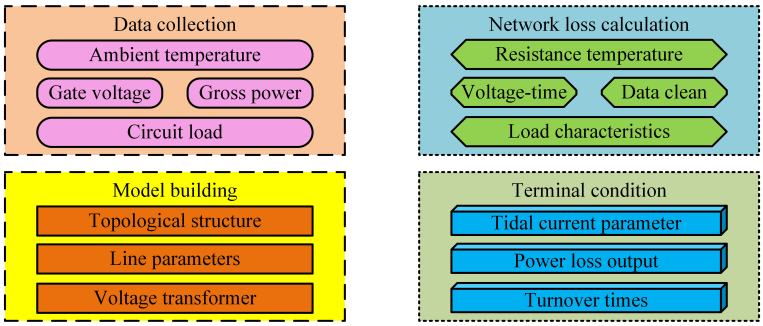


Figure 5: Calculation flow of distribution network loss

Figure 5 is the flow chart of network loss calculation using the DWKM-LSTMIA algorithm. As can be seen from Figure 5, the research first collects the relevant data of the distribution network, including power supply load, line parameters, and transformer parameters, and cleans the collected data. Then, according to the topological structure of distribution network, the calculation model of the distribution network loss is established. This model is based on the power flow calculation model.

The research uses the power flow calculation method to calculate the circuit parameters of each node in the distribution network. In this process, the transfer method of circuit parameters is related to equation (11).

$$\max \partial \frac{\Delta Vo}{\partial \Delta la} = \begin{bmatrix} \partial \frac{\Delta Uk_l}{\partial \varpi_1} & \partial \frac{\Delta Uk_l}{\partial \varpi_2} \\ \partial \frac{\Delta Uk_l}{\partial \varpi_3} & \partial \frac{\Delta Uk_l}{\partial \varpi_4} \end{bmatrix} \quad (11) \tag{11}$$

Equation (11) is a calculation method of voltage correction. In this equation, the accurate state values of the network in the adjacent time are respectively expressed in $\varpi_1, \sigma_2, \omega_3, \sigma_4$, and the line voltage and current between them are recorded as $\Delta Vo, \Delta Ia$. This step can determine the distribution of power flow in the power grid, and distribute the total loss in the power grid according to the calculation results. Research and analysis of the obtained network loss data can change the circuit resistance, as shown in Equation (12).

$$E_r = \sum_{i=1}^{20} E_i * [\theta * (Th - 20) + 1] \tag{12}$$

In Equation (12), the electrical properties of wire materials are recorded as θ, the ambient temperature is expressed as Th, and the resistance value at this temperature is recorded as E_{rr}, E_i represents the corrected resistance value. This method can compare the network loss data in different time periods, and then find out the problems in the operation of the power grid. After the implementation of optimization measures, the network loss is re-calculated, and the optimization effect is evaluated. When the effect is not ideal, further adjustment measures will be introduced, as shown in Figure 6.

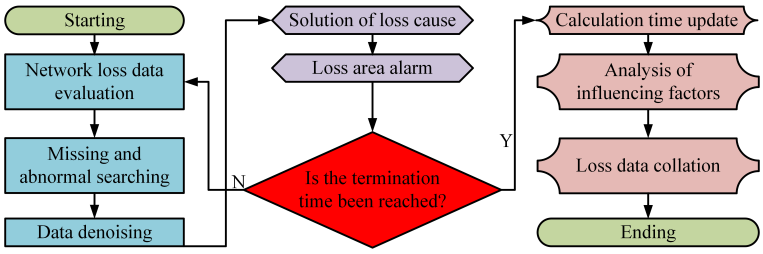


Figure 6: Process diagram of adjustment measures and methods when the result is not ideal

Figure 6 shows the processing method when the distribution network loss data has errors. As can be seen from Figure 6, the research first verifies the distribution network loss data and uses data validation rules to detect outliers. Then the missing values are filled with data alarm, and the data are made more stable through data resampling and smoothing [27]. For outliers in network loss data, a quality control mechanism is established. This method achieves the purpose of correcting data errors by regularly checking the quality of data. The equation for the inspection method is shown in Equation (13).

$$\begin{cases} Da_0 = \sum_{r=1}^{R} Bi_r{}^2 E_r \\ Bi_r = \left(\frac{\Delta U}{\Delta U k_l}\right)^r * Po_0 + E_i * \left(\frac{Si(\vartheta, \varsigma)}{\Delta U}\right)^2 \end{cases} \tag{13}$$

In Equation (13), the randomly selected data is represented by Da_0, the theoretical data by Bi_r, and the corresponding circuit power is recorded as $Po_0.Si$ is the bias coefficient of circuit connection, which is related to the value of Voltage Interaction (VI). Voltage interaction refers to the mutual complementation of transformer equipment level conversion in the process of interconnection. It plays an important role in the power system and needs to be monitored and controlled, as shown in Equation (14).

$$\mathrm{Si}(\vartheta, \varsigma) = \max_{\vartheta, \varsigma \to n^0} \frac{\int \sin \vartheta^* \ln(\tau_d) / \tau_0 d\vartheta d\tau}{\ln \min \cos \varsigma} \tag{14}$$

In Equation (14), the transformer detection coefficient is recorded as ϑ, ς, the sample size with this feature is expressed by τ_d, and the initial state of these features is recorded as τ_0. This method can be used to monitor the circuit parameters of different levels, so as to understand the operation

status of the power system and take timely measures to adjust and control. It plays an important role in the process of circuit energy transmission, and can efficiently mine association rules for large-scale power grid data sets [28]. The proposed DWKM-LSTMIA algorithm can deal with the long-term dependence in time series data, and then better capture the time correlation in the data. DWKM-LSTMIA can automatically learn the feature representation of data, without manually designing features, which effectively reduces the workload of feature engineering. By dividing the data set into different clusters, DWKM-LSTMIA can help discover the hidden rules in the data, thus providing a deeper understanding of the data.

4 Result and discussion

In order to explore the application effect of the DWKM-LSTMIA algorithm in large-scale network loss data set association rule mining, this research conducted an experiment on the Netloss data set. This dataset contained 1436 power grid lines with different voltages, including forest land, desert, and other special terrain. Rule mining was conducted on the Netloss dataset to verify the efficiency of the proposed algorithm.

4.1 DWKM-LSTMIA algorithm performance verification for power data set

In order to make rational use of the limited data in the Netloss dataset, this research divided it into two groups according to the ratio of 2:3 and conducted algorithm learning and experimental verification respectively. The equipment screening and parameter determination in the experiment are shown in Table 1. In order to fully explain the data, the study first used enough labels and legends, in which labels were used to identify different data points in the data chart. For charts with multiple data series, the legend was used to understand the meaning of each series. Both of them can explain the name of each series, and the data marker can visually show the specific value of each data point.

Table 1: Equipment selection and parameter determination in performance verification experiment of DWKM-LSTMIA algorithm

Equipment selection	Parameter determination
Data set	Netloss
Language	Easy Chinese
Master client	Intel Yeon E5-2023
Algorithm working time	15:00:54
Memory of graphics card	504G
Operating system	Windows XP
Distribution network topography	Woodland, desert, ocean, and swamp
Chord length of distribution network	28.64km
Model volume quantization	4.01
Tanh	QZ8200
Execution method	Matlab R2021a

The research verified the performance of the algorithm DWKM-LSTMIA after setting the parameters according to Table 1, and compared it with the experimental results of the Extreme Gradient Boosting (XGBoost) algorithm, the Generalized Predictive Control (GPC) algorithm, and LSTM, and obtained the superior performance comparison of the DWKM-LSTMIA algorithm. Selecting parameter values and model architecture was an important decision-making process in the DWKM-LSTMIA algorithm, which directly affected the performance and generalization ability of the model. When selecting parameter values and model architecture, the research first considered the characteristics and background of the data. Different data sets had different data sizes and noise levels, and these characteristics were used to select appropriate model parameter values. Then, different model architectures and parameter settings were considered based on different problem types, which limited the computing resources and time for the rule mining problem of electrical data sets. In addition, the knowledge of domain experts also affected the choice of research. When choosing model architecture, hyperparameter tuning refers to those parameters that cannot be learned automatically through

the training process. By trying different parameter values and using cross-validation technology to evaluate the performance of the model, the study can choose the best parameter values and model architecture. The experimental results are shown in Figure 7.

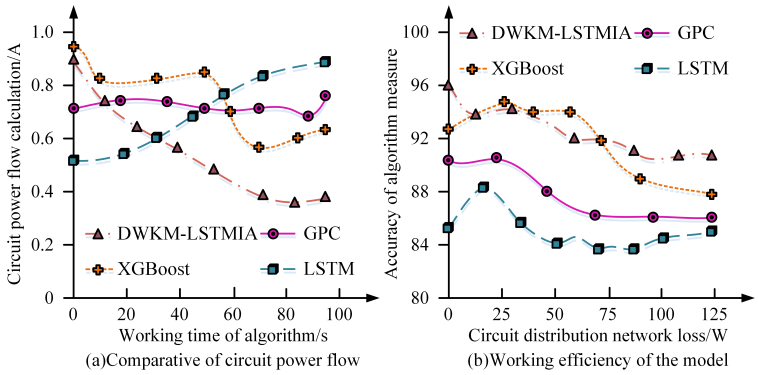


Figure 7: Comparison diagram of internal performance test of model in foundation deformation of ancient buildings

Figure 7 shows a comparative experiment on the internal performance of the four models in data rule mining. In Figure 7 (a), as the working time of the models increased, the circuit power flow of the four models decreased. Among them, the tidal current of the DWKM-LSTMIA model was the lowest, 0.37A, and the results of XGBoost, GPC, and LSTM were 0.64, 0.79, and 0.82A respectively. It showed that the DWKM-LSTMIA model had the best effect on circuit power flow control. It can be seen from Figure 7 (b) that the measurement accuracy of the model was inversely proportional to the power grid loss. Their accuracy rates were 91.7%, 88.1%, 87.6%, and 84.4% respectively, which showed that the model was the best for processing large-scale data sets. In order to verify the stability and robustness of the DWKM-LSTMIA model, experiments were conducted on different terrains. The experimental results are shown in Figure 8.

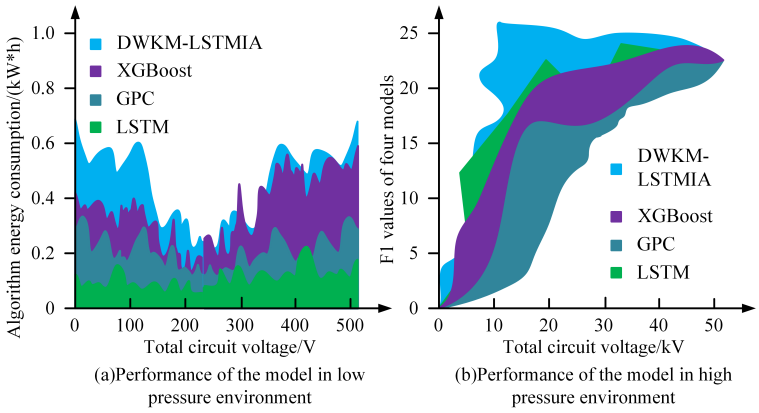


Figure 8: Comparison chart of experimental results of data fluctuation of four models

In Figure 8, the energy consumption of the four models was compared with the F1 value. It can be seen from Figure 8 that with the continuous increase of voltage, the energy consumption and F1 value of the model showed an upward trend, and the performance of the DWKM-LSTMIA model was optimal. When the circuit voltage reached 253V, the energy consumption of the model was the lowest, 0.15 kW * h. In the high voltage test, the F1 value of the fusion model performed best when the voltage value was 14kV, which was 2.4, indicating that the model proposed in the study performed better in terms of internal performance. However, this experiment can only explain the performance of the model in rule mining of data sets. In order to verify the actual performance of the model, the research carried out experimental verification.

4.2 DWKM-LSTMIA algorithm efficient experiment in large-scale data set

In order to improve the application efficiency of the DWKM-LSTMIA algorithm in association with rule mining of large-scale data sets, this research conducted experiments on data processing speed and voltage amplitude. The experimental results are shown in Figure 9.

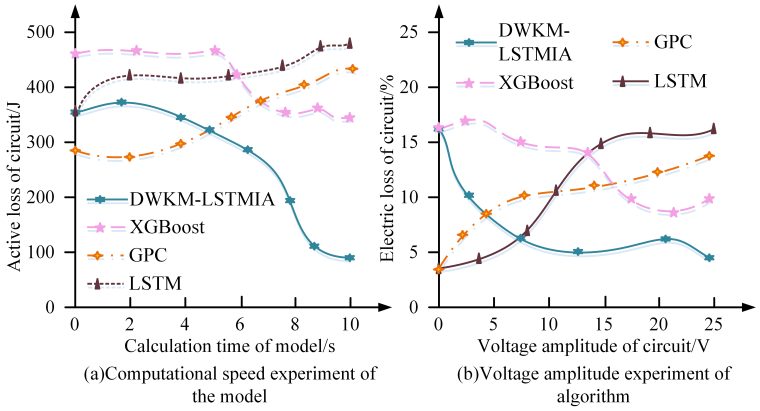


Figure 9: Experiment of data processing speed of DWKM-LSTMIA algorithm

Figure 9 is an experiment on the active power loss and network loss of the circuit. In Figure 9 (a), the active power of the four models showed an upward trend with the increase in the time for grid data calculation. Among them, the data of the DWKM-LSTMIA model was the lowest, 97J, and the experimental data of the other three models were 312, 405, and 443J, respectively, which showed that the fusion algorithm model had the highest cost performance and effectively saved economic performance. The experimental results in Figure 9 (b) showed that the line network losses of DWKM-LSTMIA, XGBoost, GPC, and LSTM were 4.7%, 8.1%, 12.5%, and 14.9% respectively. However, this test can only show the effectiveness of the fusion model. In order to verify its advantages, the study conducted experiments on the laying of power grids in different terrains, as shown in Figure 10.

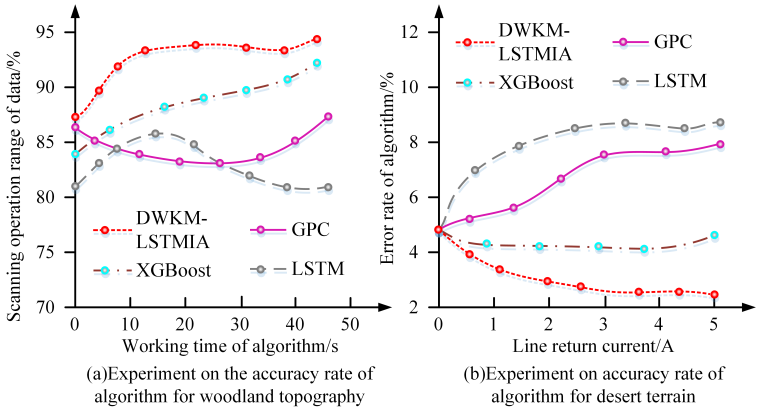


Figure 10: Experiment on the superiority of DWKM-LSTMIA algorithm

Figure 10 is an experiment on the superiority of the DWKM-LSTMIA algorithm. In Figure 10 (a), the coverage range of the four algorithms for grid data was proportional to the running time of the algorithm. The data range of the fusion model was the largest, 94.6%, which showed that this algorithm processed power grid data well. It can be seen from Figure 10 (b) that with the increase of return current in the circuit, the data error rates of the four models showed an upward trend. The data errors of DWKM-LSTMIA, XGBoost, GPC, and LSTM were 2.27%, 4.35%, 7.64%, and 8.17% respectively, of which the error of DWKM-LSTMIA algorithm was the smallest. However, this experiment alone cannot explain the universality of the fusion model, so the study optimized the model according to the evaluation results shown in Figure 10. In order to test the benchmark data set and indicators more thoroughly, different combinations of superparameters were studied and tried, and

the performance of the model was gradually improved. This method can evaluate the performance of the model more accurately. At the same time, the data set was divided into several subsets, one of which was used as the test set and the rest as the training set, and 40 experiments were carried out. The experimental results are shown in Figure 11.

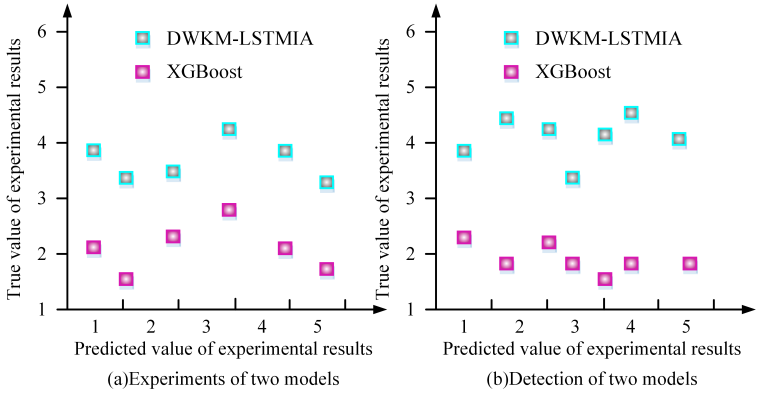


Figure 11: Forty experimental results of DWKM-LSTMIA algorithm

Figure 11 is a comprehensive experiment on the proposed fusion model. In the 40 comparative experiments in Figure 11, the linear fit of the DWKM-LSTMIA algorithm and XGBoost were 0.9993 and 0.9824 respectively. The experimental results showed that the proposed fusion algorithm effectively calculated the power grid data, and then had substantial and efficient mining of association rules for large-scale data sets [29].

5 Conclusion

With the development of electrical equipment, the amount of data generated by it increases rapidly. In order to speed up the rule mining of data, this research establishes a data warehouse in KM for efficient data analysis and combines the LSTM improved by IA with DWKM to generate a fusion algorithm. The experiment is carried out on the Netloss dataset and compared with the experimental results of XGBoost, GPC, and LSTM models to verify the effectiveness and superiority of the algorithm. The circuit power flows of the four algorithms were 0.37, 0.64, 0.79, and 0.82A respectively, which showed that the DWKM-LSTMIA algorithm had the best performance for circuit power flow control. The measurement accuracy of this algorithm was 91.7%, which was the highest among the four algorithms. For the robustness test of the algorithm, the energy consumptions of the four algorithms in the low-voltage environment were 0.15, 0.23, 0.37, and 0.52 kW * h, respectively. The F1 values of the model under a high-pressure environment were 2.4, 3.7, 6.2, and 6.5 respectively, indicating that the algorithm effectively coped with the extreme external environment. With the increase in the calculation time of power grid data, the active power loss of the DWKM-LSTMIA algorithm was the lowest, which was 97J. The experimental data of the other three models were 312, 405, and 443J respectively, which showed that the fusion algorithm created a lot of economic benefits. The line network losses of DWKM-LSTMIA, XGBoost, GPC, and LSTM were 4.7%, 8.1%, 12.5%, and 14.9% respectively, which showed that the circuit control performance of the fusion algorithm was superior. Under the same internal environment, the data operation ranges of the four algorithms were 94.6%, 91.2%, 87.5%, and 79.6% respectively. The data error of DWKM-LSTMIA was 2.27%, which was the best among the four algorithms, indicating that this algorithm had the best processing effect on power grid data. In the extensive test of the research, the linear fitting degrees of the DWKM-LSTMIA algorithm and XGBoost were 0.9993 and 0.9824 respectively. The experimental results showed that the proposed fusion algorithm had the strongest ability to mine rules from large-scale datasets, and performed well in robustness and stability. The algorithm has a high computational complexity in operation, and there are limitations in parameter tuning technology when extracting features from small-scale data sets. At the same time, the problems of over-fitting and over-generalization are improved only for large-scale data sets, and the reliability of the algorithm is biased in small-scale data

sets. Although the algorithm performs well on large-scale data sets, rule mining in small-scale data sets is equally important, thus increasing the diversity and richness of data sets, which will be gradually carried out in future research. In conclusion, this work introduced a tailored fusion algorithm for efficient association rule mining in large-scale power grid datasets. Techniques from clustering, neural networks, and association rule learning are synergistically combined. Extensive experiments highlight advantages over conventional methods in critical performance metrics like power flow, accuracy, active power loss, and network loss. These results showcase the promise of the approach for enabling scalable, real-time analytics in smart power systems. However, limitations exist in broader validation across multiple public benchmark datasets. Future work can focus on expanding testing, enhancing algorithm explainability, and exploring additional techniques like ensemble learning. This research provides key insights into advanced analytics for next-generation power grids with far-reaching societal and economic impacts.

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Et-taleby, A.; Chaibi, Y.; Boussetta, M. (2022). A novel fault detection technique for PV systems based on the KM algorithm, coded wireless Orthogonal Frequency Division Multiplexing, and thermal image processing techniques. *Solar Energy*, 237(May), 365-376, 2022.

[2] Oslund, S.; Washington, C.; So, A.; Chen, T.; Ji, H. (2022). Multiview Robust Adversarial Stickers for Arbitrary Objects in the Physical World. *Journal of Computational and Cognitive Engineering*, 1(4), 152-158, 2022.

[3] Zhou, Z.; Li, J.; Tu, J. (2021). Clustering of nasopharyngeal carcinoma intensity-modulated radiation therapy plans based on KM algorithm and geometrical features. *International Journal of Radiation Research*, 19(1), 13-21, 2021.

[4] Zhang, C.; Zhao, Y.; Zhou, Y.; Zhang, X.; Li, T. (2022). A real-time abnormal operation pattern detection method for building energy systems based on association rule bases. *Building Simulation*, 15(1), 69-81, 2022.

[5] Zhang, Q.; Geng, G.; Tu, Q. (2023). Association mining-based method for enterprise's technological innovation intelligent decision making under big data, *International Journal of ComputersCommunications&Control*, 18(2), 5241, 2023.https://doi.org/10.15837/ijccc.2023.2.5241

[6] Mao, Y., Liu, S. & Gong, D. (2023). A Text Mining and Ensemble Learning Based Approach for Credit Risk Prediction. *Tehnički vjesnik*, 30 (1), 138-147. https://doi.org/10.17559/TV-20220623113041

[7] Yang, Y.;, Tian, N.; Wang, Y.; Yuan. (2022). A Parallel FP-Growth Mining Algorithm with LoadBalancing Constraints for Traffic Crash Data, *International Journal of Computers Communications&Control*, 17(4), 4806, 2022.https://doi.org/10.15837/ijccc.2022.4.4806

[8] Meesala, S. R.; Subramanian, S. (2022). Feature-based opinion analysis on social media tweets with association rule mining and multi-objective evolutionary algorithms. *Concurrency and Computation: Practice and Experience*, 34(3), 1-25, 2022.

[9] Kota, V.; Munisamy, S. (2022). High accuracy offering attention mechanisms based deep learning approach using CNN/bi-LSTM for sentiment analysis. *International Journal of Intelligent Computing and Cybernetics*, 15(1), 61-74, 2022.

[10] Burlăcioiu, C., Boboc, C., Mirea, B., Dragne, I. (2023), Text Mining In Business. A Study of Romanian Client's Perception with Respect to Using Telecommunication and Energy Apps. *Economic Computation and Economic Cybernetics Studies and Research*, 57(1), pp. 221-234, DOI:10.24818/18423264/57.1.23.14

[11] Filali,A. E., Lahmer, E. H. B., & Filali S. E.(2022). Machine Learning techniques for Supply Chain Management: A Systematic Literature Review. *Journal of System and Management Sciences*, 12(2), 79-136, 2022.

[12] Liu, D.; Yang, F.; Liu, S. (2021). Estimating wheat fractional vegetation cover using a density peak KM algorithm based on hyperspectral image data. *Journal of Integrative Agriculture*, 20(11), 2880-2891, 2021.

[13] Antonello, F.; Baraldi, P.; Shokry, A. (2021). A novel association rule mining method for the identification of rare functional dependencies in complex technical infrastructures from alarm data. *Expert Systems with Applications*, 170(May), 114560, 2021.

[14] Zhang, J. (2020). Interaction design research based on large data rule mining and blockchain communication technology. Soft Computing, 24(21), 16593-16604, 2020.

[15] Ling, G.; Yu, C.; Wei, L. (2022). Administration Rule of Hyperlipidemic Acute Pancreatitis Based on Data Mining. *World Journal of Integrated Traditional and Western Medicine*, 8(2), 31-40, 2022.

[16] Peng, F., Sun, Y., Chen, Z. & Gao, J. (2023). An Improved Apriori Algorithm for Association Rule Mining in Employability Analysis. *Tehnički vjesnik*, 30 (5), 1435-1442, 2023.

[17] Okada, D.; Nakamura, N.; Setoh, K. (2021). Genome-wide association study of individual differences of human lymphocyte profiles using large-scale cytometry data. *Journal of Human Genetics*, 66(6), 557-567, 2021.

[18] Guo, Y.; Mustafaoglu, Z.; Koundal, D. (2022). Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. *Journal of Computational and Cognitive Engineering*, 2(1), 5–9, 2022.

[19] Syarofina, S.; Bustamam, A.; Yanuar, A. (2021). The distance function approach on the Mini Batch KM algorithm for the DPP-4 inhibitors on the discovery of type 2 diabetes drugs. *Procedia Computer Science*, 179, 127-134, 2021.

[20] Lobo, J.; Bettencourt, L.; Smith, M.; Ortman, S. (2020). Settlement scaling theory: Bridging the study of ancient and contemporary urban systems. *Urban Studies*, 57(4), 731-747, 2020.

[21] Fang, W.; Zhuo, W.; Song, Y. (2023). Δfree-LSTM: An error distribution free deep learning for short-term traffic flow forecasting. *Neurocomputing*, 526(May. 13), 180-190, 2023.

[22] Murti, Y. S. & Naveen, P.(2023). Machine Learning Algorithms for Phishing Email Detection. *Journal of Logistics, Informatics and Service Science*,10(2), 249-261, 2023.

[23] Ünver, M.; Olgun, M.; Türkarslan, E. (2022). Cosine and cotangent similarity measures based on Choquet integral for Spherical fuzzy sets and applications to pattern recognition. *Journal of Computational and Cognitive Engineering*, 1(1), 21-31, 2022.

[24] Kesiman, M.; Dermawan, K. (2021). AKSALont: Automatic transliteration application for Balinese palm leaf manuscripts with LSTM Model. *Jurnal Teknologi Dan Sistem Komputer*, 9(3), 142-149, 2021.

[25] Guo, C. (2020). The evaluation model of reconstruction effect of ancient villages under the influence of epidemic situation based on big data. *Journal of Intelligent & Fuzzy Systems*, 39(6), 8813-8821, 2020.

[26] Chun, Y. H., & Cho, M. K.(2022).An Empirical Study of Intelligent Security Analysis Methods Utilizing Big Data. *Journal of Logistics, Informatics and Service Science*, 9(1), 26-35, 2022.

[27] Meng, X.; Xiong, Y.; Shao, F. (2020). A large-scale benchmark data set for evaluating pan-sharpening performance: Overview and implementation. *IEEE Geoscience and Remote Sensing Magazine*, 9(1), 18-52, 2020.

[28] Bottin, M.; Peyre, G.; Vargas, C. (2020). Phytosociological data and herbarium collections show congruent large-scale patterns but differ in their local descriptions of community composition. *Journal of Vegetation Science*, 31(1), 208-219, 2020.

[29] Wang, S., Song, A. & Qian, Y. (2023). Predicting Smart Cities' Electricity Demands Using K-Means Clustering Algorithm in Smart Grid. *Computer Science and Information Systems*, 20(2), 657–678, 2023.

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Sun, Q.Q. (2024). Enhancing Power Grid Data Analysis with Fusion Algorithms for Efficient Association Rule Mining in Large-Scale Datasets, *International Journal of Computers Communications & Control*, 19(3), 6232, 2024.
https://doi.org/10.15837/ijccc.2024.3.6232