Research on Fault-Environment Association Rules of Distribution Network Based on Improved Apriori Algorithm

Maoran Xiao¹, Yuanyuan Sun¹* and Kejun Li¹

¹ School of electrical engineering, Shandong University, Jinan, Shandong Province, 250061, China
*Corresponding author’s e-mail: sunyy@sdu.edu.cn

Abstract. As the urban power grids gradually enter the high reliability level, the distribution network risk early warning becomes the key to further improve the reliability level. Distribution network faults have the characteristics of strong randomness and weak causality, and conventional methods are difficult to find their laws. The idea of data mining is introduced in this paper. Based on the analysis of various types of fault data, the improved Apriori algorithm is used to mine the strong correlation rules of various influencing factors in the distribution network, and the fault-environment pattern recognition library of distribution network is established to lay the foundation for the early warning of distribution network operation risk.

1. Introduction

With the rapid development of urban distribution networks, China’s urban power grid has gradually entered a stage of high reliability. In 2018, 24 cities across the country achieved a high power supply reliability level of 99.99%. Thanks to the promotion of comprehensive power outage management and live working, the key factors affecting the reliability of urban distribution networks are shifting from planned power outages to unplanned power outages. By establishing a distribution pattern of the distribution network fault-environment scene, the targeted early warning of the potential accident of the distribution network will help to further improve the reliability of the distribution network.

With the gradual promotion and application of distribution automation equipment and transformation, the data conditions of the distribution network have been greatly improved, which provides the possibility to find out the rules between different characteristics of equipment fault events [1]. However, the current research mainly focuses on the prediction of distribution network faults caused by weather factors, animal factors and plant factors, but rarely analyses the correlation between the factors affecting distribution network faults. Reference [2] proposed a fault blackout prediction method related to storm weather, using historical blackout data and weather data to establish an empirical model, and evaluating the effectiveness of the prediction model through actual storm data. Reference [3-4] using statistics model predicts distribution network failure caused by severe weather such as hurricane and typhoon. Reference [5] considered the factors affecting vegetation growth, and predicted the distribution network failure caused by vegetation growth based on the characteristics of temperature, precipitation and annual trimming times of vegetation near the feeder. Reference [6-7] predict weather-induced distribution network failures based on wind speed and lightning data.

Based on the current research, this paper proposes an association rule mining algorithm based on improved Apriori algorithm, and uses cloud transform algorithm to discretize the continuous quantity. By establishing the dimension matrix to reduce the amount of scanning and space required for excavation, this paper completes the mining network fault events and the strong association rules mining.
between various factors, and finally establish the fault pattern of the distribution network fault environment based on historical fault database.

2. Data pre-processing of association rule mining

2.1. Cloud theory

Cloud is a model of uncertainty conversion between a qualitative concept expressed by a linguistic value and its quantitative representation, which constitutes a mapping between qualitative and quantitative. In order to solve the problem that traditional association rule mining algorithm can’t deal with continuous data, this paper transforms accurate data into appropriate qualitative concepts through cloud model to realize the discrete data.

2.2. Basic concept of cloud transformation

Supposing $U$ is a quantitative domain expressed by exact numerical values, and $C$ is a qualitative concept on the domain. If the quantitative value is $x \in U$, and $x$ is a random implementation of the qualitative concept, the degree of certainty is a stochastic random number, and the distribution of $C$ from the domain $U$ to the interval $[0,1]$ in the number field is called the cloud.

Given an irregular spatial data distribution $X$ in the universe, according to the actual distribution of the frequency of its attribute values, automatically generate a superposition of several clouds $C(E_a, E_m, E_o)$ of different sizes, each cloud representing a discrete concept. This method of extracting discrete and qualitative concepts from continuous data distribution and realizing the data softening process is called cloud transformation. The formula is as shown in (1).

$$f(x) \rightarrow \sum_{i=1}^{n}[a_iC(E_{a_i}, E_{m_i}, E_{o_i})] + \varepsilon$$

In formula (1), $f(x)$ is the probability distribution function of the data; $a_i$ is the amplitude coefficient; $n$ is the number of cloud models extracted from the continuous data; $\varepsilon$ is the error threshold.

According to the contribution of the data value of the high frequency to the qualitative concept, the cloud transform takes the local maximum point in the data probability distribution as the convergence center of the data, and forms a corresponding number of cloud models for the mathematical expectation.

2.3. Data preprocessing of distribution network fault characteristics

For each distribution network fault event, the characteristics of these fault times can be obtained through the operation and maintenance personnel log, including time, section, meteorological factors, fault causes, and fault consequences. As time goes on, these fault event records will form a huge database.

For each distribution network fault time record, there are more than 10 kinds of dimension information. When mining multi-dimensional association rules, it is necessary to filter the dimensions of participation mining. Therefore, in the data of multiple dimensions, it is necessary to select data that the user cares about, the data record is complete, and the fault feature has direct or indirect connection to perform association rule mining. This paper selects the following six typical data to form a description of the fault scenario and mines the association rules as shown in table 1.

| Feature Attribute | Feature Dimension | Feature Description |
|-------------------|-------------------|---------------------|
| Space             | fault section     | distribution area by geographic location |
| Time              | date              | the date of the fault occurred during the year |
|                   | time              | the time of the fault occurred during the day |
| Meteorological    | temperature       | temperature recorded when the fault occurred sunny, cloudy, light rain, moderate rain, heavy rain, snow, heavy snow thunderstorm | yes (thunder shower, lightning warning), no |
When discretizing the "date", firstly convert the "date" into a continuous number of numbers that change between "1~365" in years and "days", and call the cloud transform algorithm to obtain the characterization. Discrete cloud models with different qualitative concepts. After the qualitative concept is divided, the stochastic decision method is used to determine which concept each continuous data belongs to. When the "temperature" is discretized, the continuous digital quantity whose value changes between "-10~40" is obtained by the discrete cloud model which characterizes different qualitative concepts. After the qualitative concept is divided, the stochastic decision method is used to determine which concept each continuous data belongs to. When the wind is discretized, the value changes between "1~10", and the cloud change process is the same as before.

## 3. Fault-Environment association rule mining based on improved Apriori algorithm

### 3.1. Classical Apriori algorithm

Suppose \( I = \{I_1, I_2, \ldots, I_n\} \) is a collection of items. Given a transaction database \( D \), where each transaction is a non-empty subset of \( I \), each transaction corresponds to a unique identifier TID (Transaction ID). For item set \( X_i \in I \), if \( X_i \in T \), the transaction is called \( X \). If there are \( k \) items in \( X_i \), it is also called \( X \) as the \( k \)-item set. Each of the transactions can contain only one-dimensional attribute \( I \), or can contain multi-dimensional attributes[9].

The association rule is an expression of \( X_i \rightarrow X_j \), and \( X_i \cap X_j = \emptyset \). If you need to mine valuable association rules in the transaction database, you need to give 2 thresholds, namely minimum support \( (S) \) and minimum confidence \( (C) \). The former reflects the minimum requirements that a set of data needs to meet in a statistical sense, and the latter reflects the user's minimum confidence in the association rules. The two formulas are as follows:

\[
S(X \rightarrow Y) = \frac{\text{count}(X \cup Y)}{\text{count}(T)} \tag{2}
\]

\[
C(X \rightarrow Y) = \frac{\text{count}(X \cup Y)}{\text{count}(Y)} \tag{3}
\]

The classical Apriori algorithm generates a frequent item set by layer-by-layer iteration, that is, generates a frequent k-item set using a known frequent (k-1) item set, and rejects a candidate set whose support is less than the minimum support degree threshold. However, the Apriori algorithm rescans the database every time it is mined. At this time, some transactions have no effect on the generation of frequent itemset. It is necessary for the algorithm to reduce the ineffective transactions in the database.

### 3.2. Improved Apriori algorithm

On the basis of the Apriori algorithm, in order to further speed up the calculation, the problem of excessive space occupied by the item set in \( C \) when \( k \) is small is solved. This paper adopts an improved algorithm that uses dimension matrix instead of set \( C \) generated in the calculation process.

After the discretization of the continuous data in the distribution network fault event record history database in Section 2, discrete data of 8 dimensions can be obtained, and the dimension matrix \( H \) is established according to the 8-dimension attributes, as shown in the following equation.

\[
M = \begin{bmatrix}
1 & A_1 & B_1 & C_1 & D_1 & E_1 & F_1 & G_1 & H_1 & n_1 \\
2 & A_2 & B_2 & C_2 & D_2 & E_2 & F_2 & G_2 & H_2 & n_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
n & A_n & B_n & C_n & D_n & E_n & F_n & G_n & H_n & n_n
\end{bmatrix} \tag{4}
\]
In the formula (4), the first column vector represents the transaction category number, each transaction category corresponds to a distribution network fault event record. The second ninth column vectors represent discrete data of 8 dimensions. The 10th column vector represents such fault the number of statistics for the event record.

The above dimension matrix is established by: calculating the risk assessment value according to each distribution network fault event record, and then scanning the xth record in the established distribution network fault event record database D to obtain 8 records of the record. If the 8-dimensional value of a row vector in the dimension matrix represents the same discrete quantity, add 1 to the statistical value in the 10th column of the row, otherwise write a new row in the dimension matrix, and set the statistic to 1. Cycle back and forth until the entire database D is scanned.

4. Case study
A total of 2,931 distribution network fault records were collected from power supply companies in a large city, ranging from January 2016 to December 2018. With the meteorological and spatial factors, a description of the fault environment scene with integrated time, space and meteorological properties is established. Among them, 2431 faults from January 2016 to June 2018 are used as training sample sets to establish a fault environment scene pattern library and a running risk criterion database. Using 500 faults from July 2018 to December 2018 as test sets, fault risk detection and matching were performed to verify the accuracy of risk matching.

Data mining based on the improved Apriori algorithm is used to set the minimum support of the lower dimension to 0.1 and the minimum confidence is 0.7. After mining, a strong association rule database is obtained as shown in table 2.

| Sort of Rule | Strong Association Rule | Support | Confidence |
|--------------|-------------------------|---------|------------|
| 1            | {Section C, C_{B3}, C_{D6}, heavy rain, C_{H3}} → {Secondary Risk} | 0.867   | 0.891      |
| 2            | {Section E, C_{B1}, C_{C3}, C_{D1}, heavy snow, C_{H1}} → {Secondary Risk} | 0.812   | 0.781      |
| 3            | {Section B, C_{B1}, C_{C4}, C_{D7}, sunny, C_{H1}} → {Tertiary Risk} | 0.761   | 0.716      |

After the training set data is used to establish the distribution network fault environment pattern identification library, the test set is used for matching test. The input part is the environment parameter of the specified scene, and the output part is the result of the equipment failure risk value. The test results are shown in the table 3. Among them, the test set is judged as one of the first, second, and third-level risks, that is, the "fault occurs" is checked.

| Sort of Area | Number of Faults in Test Set | Precision Rate of Fault/% |
|--------------|------------------------------|---------------------------|
| 1            | 34                           | 71                        |
| 2            | 41                           | 79                        |
| 3            | 32                           | 70                        |
| 4            | 45                           | 79                        |
| 5            | 39                           | 61                        |
| 6            | 23                           | 74                        |
| 7            | 38                           | 74                        |
| 8            | 34                           | 79                        |
| 9            | 48                           | 68                        |

It can be seen that the accuracy of fault occurrence is generally between 70% and 80%. These matching results indicate that the establishment of the distribution network fault-environment pattern identification library can be used as a reference for the decision of the production and operation personnel.
5. Conclusion
The case shows that the improved Apriori algorithm can find the relationship between typical faults and environment scenes well. The typical fault environment scene pattern library has good guidance and can provide a basis for urban distribution network accident prevention and control. However, this paper does not conduct a more in-depth study on the matching method, and does not further study how to more accurately predict the potential failure of the distribution network based on the fault environment pattern recognition library. This is also the content that needs to be studied in the next step.

References
[1] J. Eno and C. W. Thompson, "Generating Synthetic Data to Match Data Mining Patterns," in *IEEE Internet Computing*, vol. 12, no. 3, pp. 78-82, May-June 2008.
[2] Zhu D, Cheng D, Broadwater R P, et al. Storm modeling for prediction of power distribution system outages[J]. *Electric Power Systems Research*, 2007, 77(8):973-979.
[3] Liu H, Davidson R A, Rosowsky D V, et al. Negative Binomial Regression of Electric Power Outages in Hurricanes[J]. *Journal of Infrastructure Systems*, 2005, 11(4):258-267.
[4] Liu H, Davidson R A, Apanasovich T V. Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms[J]. *Reliability Engineering and System Safety*, 2008, 93(6):897-912.
[5] Radmer D T, Kuntz P A, Christie R D, et al. Predicting Vegetation Related Failure Rates for Overhead Distribution Feeders[J]. *Power Engineering Review*, IEEE, 2002, 22(9):64-64.
[6] Li H, Treinish L A, Hosking J R M. A statistical model for risk management of electric outage forecasts[J]. *Ibm Journal of Research and Development*, 2010, 54(3):8:1-8:11.
[7] Zhou Y, Pahwa A, Yang S S. Modeling weather-related failures of overhead distribution lines[J]. *IEEE Transactions on Power Systems*, 2006, 21(4):1683-1690.
[8] Y. Feng, W. Wu, B. Zhang and W. Li, "Power System Operation Risk Assessment Using Credibility Theory," in *IEEE Transactions on Power Systems*, vol. 23, no. 3, pp. 1309-1318, Aug. 2008.
[9] J. M. Luna, F. Padillo, M. Pechenizkiy and S. Ventura, "Apriori Versions Based on MapReduce for Mining Frequent Patterns on Big Data," in *IEEE Transactions on Cybernetics*, vol. 48, no. 10, pp. 2851-2865, Oct. 2018.