The incidence relation research of Shanghai distribution network construction cost based on big data analytics

Kailing Chen¹, Wen gu¹, Songfeng Shi¹, Fei Fei¹, Jinxiong Yu², Yuxin Xie²* and Weijun Chen²

¹ State Grid Shanghai Electric Power Company Economic and Technology Research Institute, Xuhui District, Shanghai, 200233, China
² School of Economics and Management, North China Electric Power University, Changping District, Beijing, 102206, China

*Corresponding author’s e-mail: Yuxin_Xie_ncepu@126.com

Abstract: This paper uses the incidence relation algorithm based on big data, studies the incidence relation between the various influencing factors of the construction cost of distribution network in Shanghai, and obtains an effective incidence relation model, which provides effective guidance for distribution network engineering projects and is of great significance to the future green development of distribution network.

1. Introduction

Under the requirement of "double carbon" goal, in order to meet the needs of further economic and social development, more and more research on the cost control of distribution network project construction. And with the development of computer technology, it is also a trend to apply the incidence relation algorithm of big data to the construction cost control of distribution network projects.

In distribution network engineering, with the improvement of green economy and energy conservation and emission reduction requirements, power development also needs to achieve green environmental protection, energy-saving economy and other goals [1]. Therefore, as the most volatile part of the construction cost of the whole power grid, many studies have proposed the method of planning and control of the cost control and fine management of the project [2-3].

In terms of big data, data mining technology has been applied in the construction of intelligent distribution network, which provides a good guarantee for the reliability and economy of the system [4]. Incidence relation mining method improvements for power distribution data that take into account data timing are also evolving in today's distribution network engineering projects [5]. In the aspect of power grid cost, the application of time series analysis and cluster analysis of power grid operation cost has gradually shown a good application prospect [6]. Based on the algorithm of big data incidence relation, this paper makes a reasonable analysis of the historical data of the construction of the distribution network project in Shanghai, which is conducive to the cost control of the construction and technological transformation project of the distribution network project in Shanghai to a higher standard, and is of great significance to the green, healthy and stable development of the distribution network in the region.
2. Incidence relation model of distribution network construction cost based on big data

2.1. Principle of association algorithm

Based on big data and Apriori algorithm, this paper is based on the cost control incidence relation model of distribution network engineering construction as the most widely studied incidence relation rule mining algorithm, Apriori algorithm is the earliest proposed and the most classic incidence relation algorithm, based on the mathematical principles including support, lift, confidence, and soon. The Apriori algorithm uses iterations and layer-by-layer search methods to mine frequent item sets, and then finds correlations between target events based on the frequent item sets that are mined.

The following are the main variables and concepts in association mining:

(1) Frequent item sets:
Frequent set is the main target of data mining association relation. All kinds of commonly used association relation mining algorithms take finding frequent set as the main research direction. From the perspective of statistics, frequent set can be defined as the set whose support is greater than or equal to the minimum support. Where support refers to the occurrence frequency of a certain transaction in the whole set.

(2) Support degree
Support degree refers to the proportion of the common occurrence frequency of several correlated data in the data set to all data in the whole data set, which can be regarded as the probability of the occurrence of several correlated data in the process of studying the association relationship. For example, if we have two data \( \alpha \) and \( \beta \), their support can be expressed as:

\[
\text{Support}(\alpha, \beta) = P(\alpha \beta) = \frac{\text{number}(\alpha \beta)}{\text{num}(\text{AllSamples})}
\]  

(1)

In the formula, \( \alpha \) and \( \beta \) represent the occurrence probability of \( \alpha \) and \( \beta \) data in the whole data set respectively, and \( \text{AllSamples} \) represents the total amount of all data. In general, data with high support can only constitute frequent item sets, but data with excessively low support cannot constitute frequent sets.

(3) Confidence degree
Confidence degree is the probability that another data will appear at the same time as another data, which is simply called conditional probability in probability theory. If you want to analyze the correlation between \( \gamma \) and \( \delta \) data, then the confidence degree of \( \gamma \) to \( \delta \) can be expressed as:

\[
\text{Confidence}(\gamma \leftarrow \delta) = P(\gamma | \delta) = \frac{P(\gamma \delta)}{P(\delta)}
\]  

(2)

In the formula, \( \gamma \) and \( \delta \) respectively represent the probability of occurrence of the two kinds of data in the whole data set. If the confidence of \( \delta \) to \( \gamma \) is high, the correlation between \( \delta \) and \( \gamma \) is strong.

(4) Lift degree
For the two types of data, \( \theta \) and \( \lambda \), the lifting degree of \( \theta \) to \( \lambda \) represents the ratio of the probability that a certain condition occurs in \( \theta \) data while a corresponding condition occurs in \( \lambda \) data and the probability of such a condition occurring in \( \lambda \) data as a whole. The formula is as follows:

\[
\text{Lift}(\theta \leftarrow \lambda) = \frac{P(\lambda | \theta)}{P(\lambda)} = \frac{\text{Confidence}(\lambda \leftarrow \theta)}{P(\lambda)}
\]  

(3)

Where \( \lambda \) and \( \theta \) respectively represent the probability of occurrence of two kinds of data in the data set. The degree of elevation represents the type of association between \( \theta \) and \( \lambda \). If the degree of elevation is greater than 1, they are valid strong association rules, if the degree of elevation is less than 1, they are invalid strong association rules, and if the degree of elevation is equal to 1, they are independent.
2.2. Data pre-processing
The computational principles of big data analysis are relatively easy to understand, but require higher total amounts and quality of data. Therefore, due to the limitations of historical data on the construction costs of distribution network projects, it may be difficult to meet the requirements of big data. Therefore, in the calculation of the correlation between the indicators, this paper adopts the method of quantitative analysis of the data according to the fixed floating scale, as follows:

1. Collect data on the historical cost of distribution network construction projects and extract cost budgets, estimates and other parallel project cost data for the design phase of the project.
2. Compare the actual engineering historical cost data with the cost budget and budget estimate data of the design phase, and select a certain confidence interval by combining the cost data of each parallel project. Actual cost metrics with excessive relative budget deviations are counted as 1 and actual cost metrics with small relative budget deviations are counted as 0, and no deviation is considered.
3. generates a cost variance table of 0 and 1.

In accordance with the above method, build a set of data applicable to the Apriori algorithm for the historical data of distribution network engineering construction costs, and then dig the correlation, as shown in figure 1.

![Figure 1. Data processing](image)

2.3. The construction cost incidence relation model of distribution network engineering projects
In power engineering project construction cost index correlation model building, the Apriori algorithm using the iterative method, first of all the cost index search out a candidate sets, and then according to the need to set the corresponding support and confidence, enhance degrees, pruning away below the set goals indicators portfolio, get frequent sets a. And then connected to frequent sets a, get the candidate frequent sets b, remove the discontent set enough support and confidence, improve degree of indicator combination, get real frequent sets b, and so on, iteration, until it is unable to find a higher level of frequent sets, corresponding to the set of frequent sets k is the output of the algorithm.

When the Apriori algorithm is applied to mining association relations, the calculation steps are as follows:

1. Organize the existing data to analyze the count to form a set of candidate indicators
2. Compare the existing set of metrics with the set threshold
3. Loop operation until there is no higher order at last Frequent item sets and then output the previous order frequent item sets that appear after the previous operation.
4. The final incidence relation calculation result which can guide the construction cost of distribution network project is obtained.
3. Empirical analysis

3.1. Deviation judgment
First of all, we obtain the cost data of the construction of the massive distribution network project, and make a preliminary judgment on it according to the degree of deviation. "In this case, the 5% deviation threshold taken by the switch station equipment original price indicator is considered to be no deviation from the budget gap of less than 5% and, conversely, a deviation other than 5%." The formula is as follows:

\[ k = \begin{cases} 
1, & (i \leq 0.95j \cup i \geq 1.05j) \\
0, & (0.95j \leq i \leq 1.05j) 
\end{cases} \]  \hspace{1cm} (4)

In the formula, \( k \) represents preprocessed data as required by the Apriori algorithm, \( j \) represents the estimated cost in the budget, and \( i \) represents the cost in the actual construction of the project.

Because different indicators of different projects have different deviation limits, the judgment of whether different indicators deviate is also different. It is precisely because of the different judgment requirements of each indicator in whether or not it deviates, that the deviation threshold required by each indicator in the data pre-processing of the Apriori algorithm is also different. For nonlinear indicators such as the number of in and out circuits and the type of insulation, the method of dealing with this paper is to obtain the most common type of indicator from historical data, and then the situation in which other large different types are selected unnecessarily on the indicator is considered to be that the item deviates in that indicator. The deviation threshold for the different indicators of the switch station project is shown in Table 1.

| Level indicators      | First level index deviation threshold value | The secondary indicators | Deviation threshold value of secondary indicators |
|-----------------------|--------------------------------------------|-------------------------|-----------------------------------------------|
| Equipment purchase expense | 5%                                        | Equipment price          | 5%                                            |
| Equipment purchase expense | 5%                                        | In and out of the way back | N/A                                           |
| Equipment purchase expense | 5%                                        | Insulation measures      | N/A                                           |
| Equipment purchase expense | 5%                                        | The freight              | 5%                                            |
| The installation cost   | 5%                                        | Cost of raw materials    | 5%                                            |
| The installation cost   | 5%                                        | Construction machinery use fee | 5%   |
| The installation cost   | 5%                                        | The management fee       | 5%                                            |
| The installation cost   | 5%                                        | Artificial cost          | 5%                                            |
| Other fees             | 7%                                        | Reserve funds            | 3%                                            |
| Other fees             | 7%                                        | Material test fee        | 5%                                            |
3.2. Incidence relation analysis based on big data

Through multiple calculations, we can draw the incidence relation between various types of engineering projects and various cost control indicators, and construct a cost incidence relation model for distribution network engineering projects, taking the switch station project as an example. The calculation results of the relationship between the switch station project cost-related indicators are shown in the table 2.

![Figure 2. The calculation results of the incidence relation relationship of each index of the switching station project](image)

| Serial number | Incidence relation                                | Support | Degree of confidence | Ascension degree | Data coverage |
|---------------|---------------------------------------------------|---------|----------------------|------------------|--------------|
| 1             | Equipment expenditure -> equipment purchase expense | 0.23    | 0.625                | 1.8              | 0.461        |
| 2             | Construction machinery use fee -> Project installation cost | 0.54    | 0.71                 | 1.61             | 0.71         |
| 3             | Management fee -> Installation cost               | 0.2     | 0.605                | 1.445            | 0.45         |
| 4             | Reserve funds -> Other fees                       | 0.288   | 0.536                | 1.211            | 0.53         |
| 5             | Equipment purchase expense -> Project as a whole  | 0.27    | 0.483                | 1.191            | 0.558        |
| 6             | Installation cost -> Project as a whole           | 0.23    | 0.522                | 1.18             | 0.442        |
| 7             | Reserve funds project -> Materials testing fee    | 0.21    | 0.546                | 1.115            | 0.426        |
| 8             | Labor project -> Other costs                      | 0.12    | 0.326                | 1.103            | 0.26         |

The data coverage for each association in the table is around 50%, which meets the requirements of the quality of the association. The results show that the original price of the equipment and the insulation measures have a greater impact on the acquisition cost of the equipment of the switch station project, especially the insulation measures, which have a support level of 0.46 for the purchase cost of the equipment, and both have a higher lift. Construction machinery use fee support for installation engineering fee reached 0.54 simultaneous confidence and lift degree of 0.71 and 1.61, respectively, it can be considered that the construction machinery use fee and installation engineering fee is more closely related to the main factors. At the same time, the calculation results of the incidence relation relationship also show the incidence relation between the first-level indicators or the second-level indicators, such as the support degree, confidence degree and lift degree of the equipment to the freight charges are 0.37, 0.543, 1.61, indicating that there is also a strong incidence relation between them, which needs to be considered in cost control. For the entire switch station project, the impact of equipment acquisition costs is greater, while the equipment under the first-level index of equipment acquisition costs, the two-level indicator of the switch station overall construction costs is also very large. Therefore, equipment acquisition fee is especially the control of the original price of equipment is the focus of the construction cost control of the switch station project.

For other types of distribution network engineering projects, the model can also be applied to calculate the cost incidence relation of distribution network engineering projects.

4. Conclusions

Based on the characteristics of distribution network construction projects, this paper applies the incidence relation algorithm to the study of the cost incidence relation relationship of distribution network projects, constructs the cost incidence relation model of distribution network projects, and calculates the support, confidence and promotion between key control indicators according to the data
related to the cost of distribution network construction projects, and obtains the incidence relation between key control indicators. In the future, Shanghai distribution network project construction cost control has a guiding role, and at the same time it has a good effect on the construction of green power grids.

Acknowledgments
Thanks for the support of "Research on Construction Cost Control of Distribution Network Engineering Based on Big Data", a science and technology project of State Grid Shanghai Electric Power Company Economic and Technology Research Institute.

References
[1] Wang Wei. Green Grid Development Theory and Empirical Research. North China Electric Power University, 2015
[2] Sun Jun. Analysis of engineering cost control and effective methods in distribution network construction. Technology and Market, 2020, 27 (12): 169-170
[3] Gao Qiong, Yu Zhongwei, Yu Zhongzhong. Fine planning and management of distribution network based on multi-dimensional linkage mode. Rural Electrician, 2020, 28 (04): 6-7
[4] Li Xinxuan, Yu Xiaoyong, Yang Guoyan, Qin Liwen, Ou Shifeng. Study on 0 Intelligent Distribution Grid Situational Awareness Method based on Micro PMU Data Mining (J/OL) Electrical measurement and instrumentation, {3}, {4} {5}:1-9 (2021-07-14).
[5] Guo Cheng, Zhang Wanda, Wang Bo. Incidence relation Rule Mining Methods Based on Distribution Data. Power Science and Engineering, 2021, 37 (05): 32-39
[6] Zheng Xiaodong. Power Grid Enterprise Power Supply Cost Forecast Model based on Big Data and Its Construction. Enterprise Reform and Management, 2020, {4} (04): 179-181