Power Investment Forecast Based on Classification and Regression Tree

Hua ZHOU\(^1\), Shuang-qing LIN\(^1\), Ming-wei LI\(^1\), Xiao SHAO\(^1\), Jian-jia ZHOU\(^1\), Yan DENG\(^1\) and Qian TAO\(^2,\ast\)

\(^1\)Sichuan Electric Power Company, State Grid Corporation of China, China
\(^2\)School of Automation Engineering, University of Electronic Science and Technology of China, China
\(\ast\)Corresponding author

Keywords: Classification and regression tree, Grey correlation analysis, Infrastructure investment.

Abstract. There are many factors affecting the scale of power network infrastructure investment, and the traditional grey theory model method does not consider multiple features at the same time on the forecast results. In this paper, we model the prediction of infrastructure investment as the classification and regression tree building hierarchical tree structure. The validity and reasonableness of the model are verified by the calculation experiment of power network infrastructure investment. Compared to the grey theory model, this method with higher accuracy takes into account a number of features that are highly relevant to infrastructure investment.

Introduction

The construction of power grid is of vital importance to the national economy and the people's livelihood. In recent years, with the development of the economy, the demand for social electricity is increasing, and the investment scale of the power grid enterprises is gradually raising. To satisfy the demand of using power, it is in need of measuring the investment scale and rational arrangement of infrastructure projects [1].

At present, electric power enterprises have begun some research work on the forecast of power network investment [2,3]. However, due to the existence of many factors and complicated relationships in the calculation of power network investment, it is still in need of studying related methods. The traditional power investment forecasting method is the grey model [1]. The grey theory method predicts the amount of investment without considering the influence of features on it leading to the poor generalization ability of model.

In this paper, aiming at the problem of lacking comprehensive consideration of the impact of features in the grey model theory, we adopt the classification and regression tree method. Firstly, we choose six features according to the grey relation analysis. Then, we build a model based on the classification and regression tree, and we carry out a test to verify the model. Finally, we prove the superiority of the model based on the classification and regression tree by comparing it with the model built based on the grey model theory.

The Prediction System of Power Network Infrastructure Investment

Basic Thinking of Building a Model

At present, the time range of data provided by power companies is from 2005 to 2012. The dataset has two characteristics. One is that the dataset contains of cities and hundreds of different indicators. The other is that the dataset only contains eight years of every city being not enough for regression aimed at each city. Some of data are shown in table 1.
Table 1. Partial data provided by power companies.

| Year | GDP ($X_i$) | Population ($X_p$) | ... | Infrastructure Investment ($Y$) |
|------|-------------|--------------------|-----|-------------------------------|
| 2005 | 182         | 1007.67            | ... | 13.78                         |
| 2006 | 195         | 1170               | ... | 15.19                         |
| 2007 | 207         | 1444.08            | ... | 31.42                         |
| 2008 | 262.88      | 1734               | ... | 12.64                         |
| 2009 | 267.8       | 2001.8             | ... | 16.85                         |
| 2010 | 285.09      | 2480.9             | ... | 14.92                         |
| 2011 | 352         | 3063.32            | ... | 29.57                         |
| 2012 | 347.57      | 3790.77            | ... | 48.51                         |

Let $X=(X_1, \ldots, X_n)$ denotes the features. Infrastructure investment is noted as $Y$.

There are about one hundred features, and many of them have missing value. The traditional method tends to consider the influence of a single feature on $Y$ separately. In this paper, the classification regression method is used to build the connection between end to end. The tree structure adopts a hierarchical approach according to different features. In order to construct the model, it is necessary to select the features.

**Power Network Infrastructure Investment Model**

The process to build the classification and regression tree model consists of feature selection, tree building and tree pruning. The classification and regression tree can be applied to regression or classification [4]. In the process, the key steps are the building of classification and regression tree and feature selection.

For the feature selection, we need to choose a method, which can overcome the shortcomings of the traditional method and be used in the power investment data [5]. The grey correlation analysis method can meet this requirement in this research. The formula for calculating the correlation degree $\xi_i(k)$ is as follows. $\sigma(X_0)$ is the standard deviation of $X_0$ and $\sigma(X_i)$ is the standard deviation of $X_i$. $X_0$ is the reference list. $X_i$ is the list to be compared.

$$
\xi_i(k) = \frac{1}{1 + \frac{x_0(k+1) - x_0(k)}{\sigma(X_0)} + \frac{x_i(k+1) - x_i(k)}{\sigma(X_i)}}
$$

Because some of the features have repeated meanings or same value among themselves, and part of the features has the same meaning as the infrastructure investment, grey correlation analysis method is not enough to find reasonable features for training. Thus, we select the first 20 features with high correlation value, and then we select 6 features from the 20 features according to meaning of the features. The 6 features are GDP, population, 220 kV power supply, 110 kV power supply, power consumption of the whole society, and maximum load of the whole society.

Classification and regression tree algorithm can look for the rule from the historical data, combine the features to build the tree, and predict the value of infrastructure investment in the future. The steps of the algorithm are as follows:

1. Preprocess the historical data: Normalize the value of six features and the value of power investment.
2. Divide the training set and the test set.
3. Traversing the values of the 6 features: Assume that splT = (feature, value). For example, when feature = $X_2$, value = 0.5, which means that traversing the values of feature $X_2$ for all cities and all years. When the value of feature $X_2$ of a city is greater than 0.5, the city's data for the year were
divided into the right group, and when the value of feature $x_2$ of a city is less than or equal to 0.5, the city's data for the year were divided into the left group.

4. Calculate the squared difference of the data set obtained from step 2 and select the divided method when the squared difference is minimum.

5. Assume that $\text{ops} = (\text{tolS}, \text{tolN})$. $\text{tolS}$ is the minimum error descent value and $\text{tolN}$ is the minimum sample number. Repeat steps 2 and 3 until the error descent value is less than $\text{tolS}$ or the sample number of segmentation is less than $\text{tolN}$.

6. Figure out the mean value of the power investment value at the leaf nodes.

Though step 1 to 6, we can build the tree to predict the infrastructure investment. The prediction accuracy of the constructed classification and regression tree can be obtained by comparing the predicted value with the real value of the data sample of the verification set.

**Experimental Result**

The dataset is used to reconstruct the forecasting model of power investment based on classification and regression tree algorithm. For the sake of testing the prediction effect of the model, this section aims at the construction of the model and the forecast of the power investment.

We choose the dataset of 2010 as the test set and the rest as the training set. Part of the built classification and regression tree is shown in the figure 1. From 0 to 5 represent the six features, and the number in the oval frames are the value of the features, which correspond to the “feature” and “value” in step 3.

![Figure 1. The tree based on classification and regression tree.](image)

The forecast value of power investment in 2010 can be obtained by putting training data applied to the classification and regression tree. The result of the forecast of power investment in 2010 is shown in figure 2 and table 2.
Table 2. Table of the result of classification and regression tree model.

| City   | Real (100 million) | Predicted (100 million) | Error (%) | City   | Real (100 million) | Predicted (100 million) | Error (%) |
|-------|--------------------|--------------------------|-----------|-------|--------------------|--------------------------|-----------|
| City1 | 29.57              | 42.98                    | 45.3%     | City10| 12.23              | 12.04                    | 1.57%     |
| City2 | 18.06              | 18.26                    | 11.5%     | City11| 3.35               | 3.66                     | 9.40%     |
| City3 | 9.19               | 8.84                     | 3.82%     | City12| 4.34               | 5.58                     | 28.73%    |
| City4 | 13.04              | 12.95                    | 0.71%     | City13| 3.66               | 3.66                     | 0.13%     |
| City5 | 4.2                | 3.66                     | 12.7%     | City14| 8.71               | 3.66                     | 57.92%    |
| City6 | 5.0                | 3.66                     | 26.7%     | City15| 3.43               | 5.58                     | 62.89%    |
| City7 | 2.42               | 3.66                     | 51.0%     | City16| 6.878              | 5.58                     | 18.76%    |
| City8 | 13.20              | 3.66                     | 0.72%     | City17| 3.61               | 3.66                     | 0.01%     |
| City9 | 5.67               | 3.66                     | 0.35%     | City18| 3.60               | 2.71                     | 0.24%     |
|       |                    |                          |           | City19| 4.76               | 5.58                     | 0.17%     |

Figure 2. Predicted value of cities.

From table 2, we can figure that the number of cities with errors less than 30% between predicted value and real value is 13 meaning 68.4% of cities is well and truly predicted. The average error also indicates the model based on classification and regression tree suits the dataset well. The average value of all cities is 24.8%.

In order to verify the performance of the power investment model based on classification and regression tree, a comparative test is conducted between a power investment model based on classification and regression tree with the grey theory model. We choose GM(1,1) model of the grey theory model. We choose GM(1,1) model of the grey theory.

Table 3. The result of GM(1,1).

| City   | Error (%) | City   | Error (%) | City   | Error (%) |
|-------|-----------|-------|-----------|-------|-----------|
| City1 | 99.03     | City7 | 2.14      | City13| 168.99    |
| City2 | 38.53     | City8 | 143.24    | City14| 15.29     |
| City3 | 2.97      | City9 | 46.98     | City15| 37.77     |
| City4 | 93.48     | City10| 176.36    | City16| 52.69     |
| City5 | 0.81      | City11| 3.84      | City17| 5.49      |
| City6 | 13.79     | City12| 5.74      | City18| 46.29     |
|       |           |       |           | City19| 127.91    |

Average error: 0.569126
Table 4. Table of the result of three models.

| Year | Average error (classification and regression tree) | Average error (GM(1,1)) |
|------|---------------------------------------------------|-------------------------|
| 2010 | 0.248                                             | 0.569                   |

| Year | Number of cities with errors less than 30% (classification and regression tree) | Number of cities with errors less than 30% (GM(1,1)) |
|------|---------------------------------------------------------------------------------|--------------------------------------------------|
| 2010 | 13                                                                               | 8                                                |

By comparison, We can find that the forecasting accuracy of the power investment prediction model based on classification and regression tree algorithm is obviously better than the prediction model based on GM(1,1). The average error of the forecasting model based on classification and regression tree is less than GM(1,1) of the prediction model, and the number of cities predicted accurately is also significantly improved.

**Conclusion**

This paper first discusses the traditional method of dealing with power investment which is the grey theory model. In the model, features are not being not taken into account, and it has higher demand for the dataset. In order to solve the problems, we introduce the method in this paper to construct classification and regression tree model. Finally, through the experiment, the results indicates the classification and regression tree model has superior results to the GM(1,1) model, and the error rate of different cities between the two models is up to about 32 percentage.

**References**

[1] Hu B C A, Hu G A, Hu C H A, et al., Grid Infrastructure Investment Calculation Model Based on Gray Prediction[J]. Dianzi Keji Daxue Xuebao/journal of the University of Electronic Science & Technology of China, 2013, 42(6):890-894.

[2] Han Liu, Peng Dong, Wang Zhi-dong, et al. Constitution and application of the index system for power grid assessment[J]. Electric Power Construction, 2010, 31(11): 28-33.

[3] Cai Guang-zong, You Jian-xin, Shang Shan-shan. The establishment of the post assessment for the construction process of power grid project[J]. Shanghai Management Science, 2012, 34(6): 107-112.

[4] Yu C. Adaptive Japanese Teaching Optimization Based on Classification and Regression Tree[C]//International Conference on Robots & Intelligent System. IEEE Computer Society, 2017:15-18.

[5] Zi-jie Wei. Research and Software realization of Power Network Investment calculation method based on Grey Theory [D]. University of Electronic Science and Technology of China, 2013.