Resource Consumption Forecasting for Information System using Holt-Winters Method and Linear Regression Method

Yunrui He, Yiying-yan, Yijie-dang, Zhiyu-chen

Dispatch and Control Center, State Grid Information & Telecommunication Co., Ltd., Beijing, 100075, P.R. China

hyr52045@163.com, mango27@126.com, dangyijie@126.com, sevsevf@163.com
heyunrui@sgcc.com.cn

Abstract. Through the analysis of historical data and model training, the resource consumption forecasting of information system can estimate the utilization of resources in the future and help to make resource planning in advance. The traditional methods are usually based on the experience of operation and maintenance personnel and the preliminary statistics of subsequent plans, which is time-consuming and labor intensive, and cannot adapt to unplanned changes. In order to solve these issues, an automatic resource consumption forecasting method based on Holt-Winters and linear regression algorithm is proposed in this paper. The experimental results show that the method can well predict the periodic and incremental resource consumption changes.

1. Introduction

There are two kinds of traditional resource consumption forecasting methods: (1) the experiential method that makes resource planning at 20 times of the current consumption, implements at 3 times and deploys at 1.5 times; (2) the statistical method that calculates the expected business volume of new information system to be launched in the latest quarter and half a year. The shortcomings of the above methods are obvious: (1) deep dependence on experts’ experience; (2) serious waste of resources; (3) long period of planning; (4) extensive management.

The forecasting method of resource consumption this paper proposes can automatically predict the future trend of resources consumption according to their historical usage, and help to make better plans in advance. The main values are as follows. (1) Improving the efficiency of operation and maintenance and the accuracy of forecasting, reducing labor costs, and achieving the best return on investment ratio; (2) From passive capacity expansion to active forecasting, realizing the trend perception of resource demand and the early warning of resource shortage and purchase in advance; (3) Greatly shorten the cycle of resource planning and improve the efficiency of resource planning; (4) More refined and intuitive resource management.

2. Resource consumption forecasting methods

According to the different application scenarios, the algorithm used for resource consumption forecasting is also different. The Holt-Winters exponential smoothing algorithm[1] is used to predict the consumption of CPU and memory for the reason that the usage of CPU and memory fluctuates periodically, and the linear regression[2] and Gaussian weighted algorithm are used to predict the consumption of storage for the reason that the usage of storage increase linearly with time.
2.1. Holt-Winters exponential smoothing algorithm

There are several different forms of exponential smoothing method\cite{3}: the single exponential smoothing method is for the series without trend and seasonality, the quadratic exponential smoothing method is for the series with trend but no seasonality, and the cubic exponential smoothing method\cite{4} is for the series with trend and seasonality. "Holt winters" sometimes refers to cubic exponential smoothing method. The principle Holt-Winters algorithm based on is that the exponential smoothing value of any period is the weighted average of the actual observation value of the current period and the exponential smoothing value of the previous period. Holt-Winters algorithm has two advantages that the effect of recent observations on the predicted values was strengthened and the weights assigned to observations are flexible.

2.1.1 Single exponential smoothing method

The predictive formula of the single exponential smoothing method is as follows. \( x_t \) is the current observation value, \( x_{t+1} \) is the predictive value, and \( \alpha \) is the smoothing factor, \( 0<\alpha<1 \).

\[
x_{t+1} = \alpha \cdot x_t + (1 - \alpha) \cdot x_{t-1}
\]  

(1)

The Recurrence formula is as follows. \( s_t \) is the current exponential smoothing value, and \( s_{t-1} \) is the exponential smoothing value of the previous period.

\[
s_t = \alpha x_t + (1 - \alpha) \cdot s_{t-1}
\]  

(2)

Equation (2) can be expanded to equation (3):

\[
s_t = \alpha x_t + (1 - \alpha) x_{t-1} + \alpha(1 - \alpha)^2 x_{t-2} + \cdots + \alpha(1 - \alpha)^{t-1} x_1 + (1 - \alpha)^t s_0
\]  

(3)

The single exponential smoothing method is suitable for time series without general trend. If it is used to process the series with general trend, the smoothing value will lag behind the original data, unless the value of \( \alpha \) is close to 1, which will cause the series to be not smooth enough.

2.1.2 Quadratic exponential smoothing method

In order to overcome the shortcoming of the single exponential smoothing method, the quadratic exponential smoothing method retains the trend information by adding trend data \( b_t \) as shown below:

\[
b_t = \beta(s_t - s_{t-1}) + (1 - \beta)b_{t-1}
\]  

(4)

The recurrence formula and the predictive formula are shown in the following two equations.

\[
s_t = \alpha x_t + (1 - \alpha)(s_{t-1} + b_{t-1})
\]  

(5)

\[
x_{t+h} = s_t + h \cdot b_t
\]  

(6)

2.1.3 Cubic exponential smoothing method

On the basis of the quadratic exponential smoothing method, the cubic smoothing method takes the seasonality of time series into account which is represented by the seasonal data \( c_t \), and \( k \) is the length of period.

\[
c_t = \gamma(x_t - s_t) + (1 - \gamma)c_{t-k}
\]  

(7)

The recurrence formula is modified to the following equation (8). \( b_t \) is the same as equation (4).

\[
s_t = \alpha(x_t - c_{t-k}) + (1 - \alpha)(s_{t-1} + b_{t-1})
\]  

(8)

The predictive formula is modified to the following equation (9).
\[ x_{t+h} = s_t + h \cdot b_t + c_{t-k+h} \]  

(9)

2.2 Linear regression and Gauss weighted algorithm

Linear regression algorithm reflects the trend of consumption growth, and Gaussian weighting reflects the volatility of the growth process.

2.2.1 Linear regression algorithm

When some input and output variables are known, a prediction model can be fitted by linear regression algorithm[5], and when there are new input variables, the corresponding output variables can be predicted. The details are follows.

Suppose the input matrix is X, the output matrix is Y, and the regression coefficient matrix is W, the relationship between them is as follows.

\[ TYX W \]  

(10)

To find out the best W with the minimum error, we can compare the predictive values with the actual values, and make the square error minimized. The square error is shown below. \( x_i \) and \( y_i \) are a set of known input and output values.

\[ \sum_{i=1}^{m} (y_i - x_i^T w)^2 \]  

(11)

The formula (11) can be expressed in matrix form as follows.

\[ (Y - XW)^T (Y - XW) \]  

(12)

The least square error can be obtained by calculating the partial derivative of W in the formula (12) and making the result equal to 0. The partial derivative of W is as follows.

\[ X^T (Y - XW) = 0 \]  

(13)

The best W with the minimum error can be obtained from the equation (13).

\[ W = (X^T X)^{-1} X^T Y \]  

(14)

2.2.2 Gauss weighted optimization

The standard linear regression is a kind of unbiased estimation, which may lead to the phenomenon of under fitting. Therefore, we need to adjust the error for different points, that is to introduce bias to reduce the mean square error of prediction. The details are follows[6].

Each point near the predicted point is given a certain weight, and then the least square method is used for ordinary linear regression.

\[ w = (X^T WX)^{-1} X^T WY \]  

(15)

Through the above formula (15), for any given unknown data the corresponding regression coefficient \( w \) can be calculated, and the corresponding predicted value \( Y \) can be obtained. \( W \) is a diagonal matrix and it is a diagonal matrix. There are many ways to calculate \( W \), the most used is Gaussian Kernel.

\[ w_{ii} = \exp\left(\frac{|x_i - \bar{x}|}{-2k^2}\right) \]  

(16)

The value of \( w_{ii} \) represents the weight of the \( i \)th sample. \( k \) is a hyperparameters used to control the local range of local weighting.
3. Resource Consumption Forecasting Implementation

The forecasting method of resource consumption in this paper is divided into seven steps to implement, the following are the specific implementation steps and the Fig.1 is the framework of resource consumption forecasting implementation.

a. The data layer provides the training sample data for AI framework.

b. Through the feature framework, the sample data is processed into data for machine learning algorithm training.

c. The algorithm is used to train the resource consumption prediction algorithm in each scene, and the training results are saved in the result data.

d. The algorithm is evaluated by comparing the result data with the actual data.

e. The algorithm is optimized by algorithm selection or parameter optimization until the model evaluation meets the requirements.

f. The final algorithm model is used to predict the resource consumption of corresponding scenarios, and the prediction data is obtained to help to get a better plan.

g. The front-end visualizes the back-end data.

![Fig.1 the Framework of Resource Consumption Forecasting Implementation](image)

4. Resource consumption forecasting application

Based on the forecasting method of the resource consumption proposed in this paper, four main functions of resource consumption prediction service management, task configuration, model evaluation and result display are realized.

4.1 Task Configuration and Model Evaluation

By setting the resource consumption prediction scenario, object type, data acquisition method and other options, the historical data required for resource consumption forecasting can be obtained.

Then evaluate the model, it will show the curve of historical data and training result data and calculate the model matching degree, as shown in Fig.2.
4.2 Result Display
It can display the resource consumption forecasting results of one month, one quarter and half a year, including total usage, historical value and forecast value of usage, as shown in Fig.3.

5. Conclusion
This paper discusses two different forecasting methods for the consumption of CPU, memory and storage of information system, and puts forward the technical framework to realize the two methods. The accuracy of the two forecasting methods is verified by the experimental results which can be used in the actual environment as a support for making decision. The selection of initial value in the two methods is particularly important, which will be one of the focuses of our future research.

Acknowledgments
This work was supported by State Grid Information &Telecommunication Branch for the project environment. This work was also supported by Huawei Technologies CO., LTD for physical resources. The authors would like to thank our leaders and collaborators.

References
[1] S. Subramanian and A. Kannammal, "Real Time Non-Linear Cloud Workload Forecasting using the Holt-Winter Model," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-6.
[2] J. Wu, C. Liu, W. Cui and Y. Zhang, "Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression," 2019 IEEE International Conference on Power Data Science (ICPDS), Taizhou, China, 2019, pp. 139-142.

[3] R. S. Soni and D. Srikanth, "Inventory forecasting model using genetic programming and Holt-Winter's exponential smoothing method," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2017, pp. 2086-2091.

[4] G. Feng, L. Chen-Yu, Z. Bin and Z. Su-Qin, "Spares Consumption Combination Forecasting Based on Genetic Algorithm and Exponential Smoothing Method," 2012 Fifth International Symposium on Computational Intelligence and Design, Hangzhou, 2012, pp. 198-201.

[5] X. Dong, J. Chen, K. Zhang and H. Qian, "Privacy-Preserving Locally Weighted Linear Regression Over Encrypted Millions of Data," in IEEE Access, vol.8, pp.2247-2257, 2020.

[6] M. Alamaniotis, S. Chatzidakis and L. H. Tsoukalas, "Monthly load forecasting using kernel based Gaussian Process Regression," MedPower 2014, Athens, 2014, pp. 1-8.