International Journal of Energy
ISSN: 2957-9473 | Vol. 2, No. 1, 2023

Short-term power load forecasting based on RF-CNN-SVM

Xiaochao Liu

School of Electrical Information Engineering, Northeast Petroleum University, Daqing, Heilongjiang, China

Abstract: In order to reduce the error of short-term power load forecasting and improve its forecasting accuracy, A prediction method based on the combination of random forest (RF), convolution neural network (CNN) and support vector machine (SVM) is proposed. First, the data is preprocessed, and the RF algorithm is introduced to optimize the input variables, Then the feature is extracted through CNN, Finally, the extracted results are input into the SVM model, and the forecasting results are output to realize the load forecasting. In this paper, the power load data of Singapore is used for experimental analysis, compared with CNN-SVM model without RF algorithm, SVM model and hybrid model of convolutional neural network and long short-term memory network (CNN-LSTM), The results show that the prediction model method proposed in this paper has better prediction effect.

Keywords: Short-term power load forecasting; RF-CNN-SVM.

1. Introduction

With the further development of modern industry and technology, the power industry has been closely connected with people's daily life, among them, power load forecasting is a key technology in the power system, which can be divided into short term, medium term and long term according to the time of forecasting. Accurate load forecasting can not only guarantee the safe and reliable operation of the system, It can also reduce the waste of limited resources, benefit environmental protection, and improve economic benefits [1,2]. Power load forecasting plays an important role in the stable and safe operation of power system; However, load change has certain uncertainty, randomness and nonlinearity, and there are various factors that affect load change. Therefore, it is necessary to consider some factors related to load change to further improve the accuracy of load forecasting, so as to formulate an effective power plan to reduce power production costs and reduce power resource consumption [3].

At present, the main methods of load forecasting at home and abroad are statistical forecasting, time series forecasting and artificial intelligence forecasting. The statistical prediction method is mainly represented by Kalman filter method; The time series forecasting method uses the load data series of a period of time to predict the change trend of the subsequent load, and has the characteristics of fast calculation speed. However, this method requires the time series to have a high stability, and this method is weak in data regression [4]; Artificial intelligence prediction method has obvious advantages in nonlinear data fitting and parameter learning, and is currently the most widely used prediction method [5,6]. The accuracy of load forecasting based on fuzzy reasoning is low [7]. Back-propagation network, namely BP neural network, has good generalization ability, but its convergence speed is slow [8]. LSTM network model is suitable for processing input time series data, but its calculation time is long [9]. Considering the disadvantages of the above methods, this paper uses the RF algorithm to process the forecast data set and the combination of CNN and SVM to forecast the power load. The RF algorithm can optimize the original data set, detect the importance of features, and select good features. CNN has obvious effect in feature extraction, while SVM has more obvious advantages in processing and dealing with small samples, non-linear and high-dimensional problems.

2. Principle of load forecasting model

2.1. Random forest algorithm

Random Forest (RF) algorithm is an integrated algorithm composed of decision trees. The random forest is mainly based on the random subspace theory and self-help method. It randomly extracts the characteristics of sample data, and then generates a decision tree. Each tree will give classification results. The classification result with the most votes in all decision trees is the final result [10]. For the load training data set, randomly select subset b, and after extracting the i-th subset, you can get the out-of-bag data, that is, the un-extracted sample data, and then select the vector with dimension m from the M-dimension vector as the input variable to form a regression tree. When selecting the input variable, you need to follow the principle of minimum variance [11], and the function expression is as follows:

$$I = \min \frac{\Sigma\left(X_s - X_s^{'}\right)^2}{n} \tag{1}$$

In the formula, I is the optimal split input variable, n is the number of samples contained in the load training data set, s is the dimension of the sample, Xs and Xs， represent the value of the variable and the arithmetic mean value respectively. Multiple decision trees form a random forest, and then further consider the off-bag data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^{'}\right) \tag{2}$$

In the formula, MSE is the mean square residual, n is the number of samples of out-of-bag data, yi and yi， represent the true and predicted values of the i-th sample, respectively. The random forest algorithm uses the reduction of the mean square residual to determine the importance of the input variables. It mainly uses the off-bag data to generate the regression tree, and then obtains the mean square residual matrix. After multiple random sampling, it can generate a new regression tree, thus forming the mean square residual matrix

of the off-bag data. For the m-th input variable, the function expression of its importance value is as follows:

$$Vim = \frac{1}{k \cdot S_E}\sum_{j=1}^{k}\left(MSE_j - MSE_{kj}\right) \quad (3)$$

In the formula, Vim is the importance value of input variables, k is the number of decision trees, and SE is the standard error of k decision trees.

2.2. Convolution neural network

Convolutional Neural Networks (CNN) is a neural network that includes convolution and related operations, and is one of the more classical neural networks in artificial intelligence methods. Similar to other neural networks, CNN has a multi-layer network structure, mainly including convolution layer, pooling layer and full connection layer.

The main function of the convolution layer is to extract data features, that is, first use the convolution check to extract local features from the input matrix, and then select the activation function, generally tanh and relu and other nonlinear functions, to map the extracted features. The main role of pooling layer is to reduce the feature dimension while ensuring effective features, which can be divided into average pooling and maximum pooling. The main function of the full connection layer is to merge the extracted features, and then send the merged features to the output layer.

2.3. Support vector machine

Support Vector Machines (SVM) is a statistical learning method, proposed by Vapnik et al., which has been mainly used in the fields of speech recognition and time series prediction in recent years, and can be used to solve classification and regression problems. This method is mainly based on the principle of structural risk minimization and the theory of high-dimensional space, and transforms the solution of convex optimization problems into linear programming problems [12]. At the same time, SVM has a series of advantages, such as less adjustment parameters, fast calculation speed and high prediction accuracy, and can better handle small sample, nonlinear and high-dimensional load forecasting problems.

Unlike traditional regression methods, SVM introduces insensitive loss factor ε. When the absolute value of the difference between the predicted value and the target value is less than ε, the prediction is considered accurate and the loss can be ignored. The prediction model based on SVM can solve the following problems:

$$\min J\left(\omega, b, \xi_i, \widehat{\xi_\iota}\right) = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}\left(\xi_i + \widehat{\xi_\iota}\right)$$

$$s.t.\begin{cases} \mathrm{y}_i - \omega^T\phi(x_i) - b \le \varepsilon + \xi_i \\ -\mathrm{y}_i + \omega^T\phi(x_i) + b \le \varepsilon + \widehat{\xi_\iota} \\ \xi_i, \widehat{\xi_\iota} \ge 0, i = 1,2,\cdots,n \end{cases} \quad (4)$$

In the formula, ωTϕ(xi)+b stands for separation hyperplane, xi is the input feature vector, that is, the data feature extracted by CNN, ω is the weight vector of the separating hyperplane, b is the deviation, ξ is the relaxation variable, yi is the target value, and C is the penalty factor. If C is larger, it means that only a few points cross the separation boundary, and the model is more complex [13].

Because some data in the input space is linearly indivisible, the kernel function is introduced to map it to a higher dimensional space, so as to accurately partition the separating hyperplane and further solve the nonlinear regression problem. For the selection of kernel function, the radial basis function (RBF) with better performance can be selected, and its function expression is as follows:

$$K\left(x_i, x_j\right) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (5)$$

In the formula, K(xi， xj) is the kernel function and σ is the width of the kernel function. The final prediction model is:

$$\mathrm{y}(x) = \sum_{i=1}^{n}(\alpha_i - \widehat{\alpha_\iota})K(x, x_i) + b \quad (6)$$

In the formula, αi is the Lagrange multiplier introduced in solving the function.

3. Experimental analysis

3.1. Experimental equipment and environment

The experimental environment in this paper mainly uses Intel i5-8300H computer processor, and the graphics card is NVIDIA GeForce GTX 1050. The main programming language is python, and then the load forecasting model is built based on Tensorflow framework and Keras library. Numpy, Pandas, Scikit-lean and Matplotlib modules are also used.

3.2. Normalization of experimental data

The normalization of experimental data is to standardize the original data through linear transformation. The range of data size is [0,1]. Since different features of data represent different meanings, and there are large differences between the values, if normalization is not carried out, it will have adverse effects on the optimization of model parameters. Therefore, this paper takes the maximum and minimum values of data features as the benchmark values for normalization, and the function expression is as follows:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (7)$$

In the formula, x* is the normalized data, x is the original sample data, and xmax and xmin are the maximum and minimum values of the sample data.

3.3. Experimental evaluation index

In order to evaluate whether the forecasting results of load forecasting are accurate or not, the forecasting accuracy can be evaluated by root mean square error (RMSE) and mean absolute percentage error (MAPE), which can deeply reflect the error level between the predicted value and the true value. The function expression is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_\iota})^2} \quad (8)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \overline{x_\iota}}{x_i}\right| \times 100\% \quad (9)$$

In the formula, n is the number of samples in the test set, xi is the actual value of the load at the i-th sampling, and $\overline{x_\iota}$ is the predicted value of the load at the i-th sampling. The smaller the RMSE value and MAPE value, the higher the accuracy of load forecasting.

3.4. Model construction and parameter determination

Considering that power load is easily affected by many factors, such as time and environment, these factors need to be input as eigenvectors for load forecasting. However, in the process of load forecasting, it is beneficial for the research of load forecasting to properly reduce the feature vectors with less influence. In this paper, the random forest algorithm is used to analyze the importance of multiple feature variables that affect the load. After data preprocessing of the 8-dimensional feature vector, the more important features are

selected. The importance analysis results of the random forest algorithm for the feature vector are shown in Figure 1.

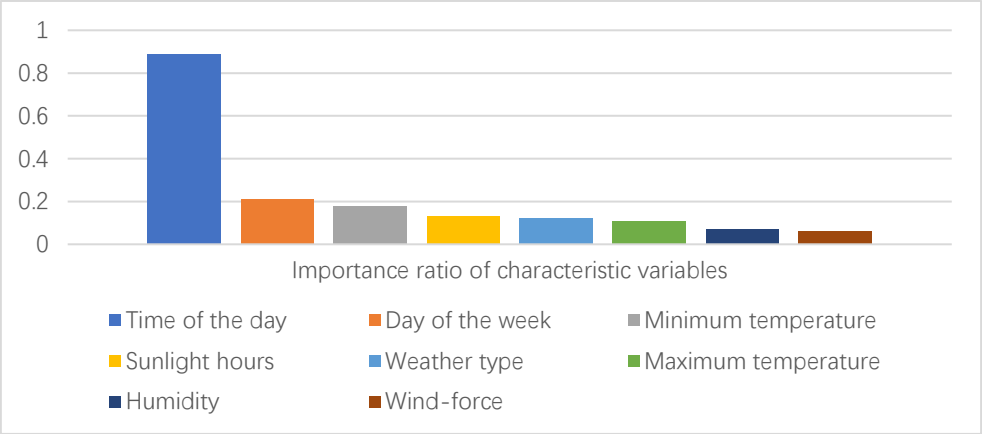


Figure 1. Analysis of importance proportion of characteristic variables

It can be seen from Figure 1 that the importance analysis of multiple characteristic variables that have influence on the load shows that the degree of influence of characteristic variables on the load is different. Considering that the characteristic variables with importance ratio less than 0.1 have less influence on the load, select to remove them, and then select the remaining characteristic variables as the input of the power load forecasting model.

The important feature vectors processed by the RF algorithm and the load data are recombined to form a new multi-dimensional feature vector, namely, the 7-dimensional feature vector, which is used as the input of the CNN model. The CNN extracts the features and sends them to the SVM model for load forecasting. The output result of the SVM is the result of the power load forecasting, thus forming a load forecasting model based on RF-CNN-SVM. Two convolution-pooling layers are used for CNN. The number of convolution cores is 32 and 64, and the size is 3×3. Convolution step size is 1×1. The activation function is ReLU. Select the maximum pooling method with the size of 2×2, in steps of 2×2. For the SVM model, RBF is selected as the kernel function of SVM, and then the penalty factor C is 100, the insensitive loss factor ε is 0.1, and the kernel width σ is 0.9.

3.5. Analysis of experimental results

This paper selects the power load data of users in Singapore from March 1, 2022 to March 30, 2022 for a total of 30 days. The load is collected every half an hour, a total of 1440 sets of data. The collected load data is divided into training set and test set, the ratio of which is 9:1, that is, the data of the first 27 days is the training set, which is used to train the built model, and the data of the last 3 days is the test set, which is used to test the prediction results of the model.

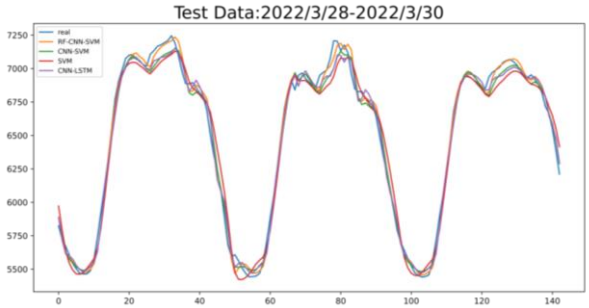


Figure 2. Comparison of prediction results of different models.

Table 1. Comparison of prediction errors of different models.

Prediction model	RF-CNN-SVM	CNN-SVM	SVM	CNN-LSTM
RMSE/MW	52.7797	59.1746	89.5615	63.7537
MAPE/%	0.6030	0.7180	1.0965	0.7586

The model in this paper is compared with CNN-SVM, SVM and CNN-LSTM, and the load forecasting accuracy of different models is tested by training different models. Since the model in this paper is RF-CNN-SVM, the relevant parameter settings of SVM model, CNN-SVM model and CNN-LSTM are consistent with the model in this paper.

After the model is trained, the data in the next three days will be predicted. The results of load forecasting are shown in Figure 2. It can be seen from Figure 2 that the load data of the next three days from March 28 to March 30 have certain regularity. The prediction result of the RF-CNN-SVM model proposed in this paper is the closest to the true value, with a good degree of fitting. The predicted value of most points coincides with or is close to the true value, with a good prediction effect.

Analyze the error level of the prediction results of different models, and the error results are shown in Table 1. It can be seen from the data in the table that the root means square error and the average absolute percentage error of the RF-CNN-SVM model proposed in this paper are the minimum. Compared with SVM and CNN-LSTM models, the error of CNN-SVM model is relatively small, which can highlight the advantages of the combined model of CNN and SVM. Compared with CNN-SVM model, the error of RF-CNN-SVM is smaller, with the root mean square error and the average absolute percentage error reduced by 0.108% and 0.115% respectively. It can be seen that the RF-CNN-SVM model with stochastic forest algorithm has higher accuracy in power load forecasting, Thus, the accuracy of the power load forecasting model is verified.

In conclusion, RF-CNN-SVM model has better effect in image fitting and prediction accuracy.

4. Conclusion

This paper proposes a short-term load forecasting method based on RF-CNN-SVM to meet the increasing demand of short-term load forecasting accuracy. It uses random forest algorithm to preprocess the data, then CNN extracts effective

features, and then inputs the extracted feature results into SVM for forecasting. This method successfully combines the different advantages of RF algorithm, CNN and SVM, and further improves the accuracy of load forecasting. Compared with CNN-SVM, CNN-LSTM and SVM models, it has better load forecasting effect.

The inadequacy of this paper is that it does not consider other factors such as electricity price and holidays that may affect the load. These factors need to be introduced and improved in the future.

References

[1] Li Xiao. Research on short-term load forecasting method based on deep learning [D]. Jiangnan University, 2021.

[2] Xu Jianjun, Huang Lida, Yan Limei, Yi Na. Insulator Self-Explosion Defect Detection Based on Hierarchical Multi-Task Deep Learning[J]. Transactions of China Electrotechnical Society, 2021,36(07):1407-1415.

[3] Limei,LIU Yongqiang,XU Jianjun,et al.Broken string diagnosis of composite insulator based on Grabcut segmentation and filler area discrimination[J].Power System Protection and Control,2021,49(22):114-119

[4] Yi, Q. Wang, L. Yan, et al., A multi-stage game model for the false data injection attack from attacker's perspective. Sustainable Energy Grids & Networks 28 (2021).

[5] Na Yi,Jianjun Xu,Limei Yan,Lin Huang. Task Optimization and Scheduling of Distributed Cyber-physical System Based on Improved Ant Colony Algorithm. Future Generation Computer Systems, 109(Aug. 2020),134-148.

[6] Yang Zhao, Jianjun Xu, Jingchun Wu. A New Method for Bad Data Identification of Oilfield Power System Based on Enhanced Gravitational Search-Fuzzy C-Means Algorithm. IEEE Transactions on Industrial Informatics. VOL. 15, NO. 11, NOVEMBER 2019 5963-5970

[7] Jing Han,Xi Wang,LiMei Yan, Aida Dahlak, et al. Modelling the performance of an SOEC by optimization of neural network with MPSO algorithm. International Journal of Hydrogen Energy, Volume 44, Issue 51, 22 October 2019, Pages 27947-27957.