Research on EA-Xgboost Hybrid Model for Building Energy Prediction

Wu Yucong* and Wang Bo
College of Computer Science, Chongqing University, Chongqing 400044, China

*1045104753@qq.com

Abstract. Building energy forecast plays an important role in Intelligent Building. Due to its non-stationarity and uncertainty, the prediction accuracy of existing methods need to be further improved. In view of this problem, propose the EA-XGBoost model, which combines Empirical Mode Decomposition (EMD), ARIMA and XGBoost model to predict building energy consumption. First, EMD is used to decompose the consumption data into multiple Intrinsic Mode Functions (IMF). Afterwards, ARIMA model is applied for each IMF to get regression result, then sum the results and calculate the residual. Taking the residual as an input feature of XGBoost, combined with other energy-related factors such as dry and wet bulb temperature, using XGBoost after Grid-Search to predict building energy consumption data. Compared with ARIMA and XGBoost model, EA-XGBoost hybrid model performs best in forecasting building energy consumption dataset which provided by the US National Renewable Energy Laboratory. The experiment shows the feasibility and effectiveness of the new model.

1. Introduction
The WMO Statement on the State of the Global Climate 2018 issued by the World Meteorological Organization in March 2019 shows that global warming is still accelerating, and record greenhouse gas concentrations have pushed global warming to increasingly dangerous levels. Economic impact is still increasing. The report also pointed out that energy consumption was the main factor that caused greenhouse effect. Building energy consumption is almost one third of the world's total energy consumption[1,2]. Building energy consumption forecast plays an important role in energy management, dispatch[3]. Data energy consumption prediction can help users estimate potential energy savings and implement effective energy management. Using artificial intelligence algorithms to build a data-driven model has certain practical value.

Many machine learning methods were applied in energy consumption, such as ANN [4], SVM [5-6] and ANN-SVM hybrid algorithms [7-8]. The experiment shows that the accuracy of ANN-SVM hybrid algorithm is higher than that of ANN and SVM, which shows that choosing the appropriate machine learning model and combining it can improve data prediction accuracy. ANN and Random Forest (RF) of ensemble learning for building energy prediction [9]. Experimental results show that ensemble learning model can be used for building energy data prediction, and the prediction effect of RF is better than ANN. The single model and ensemble model are compared in the prediction[10], and the result shows that the latter one has better accuracy and generalization ability, but the performance depends on the choice of the basic model, which needs experience and to be studied. The selection the basic model based on prior knowledge, which shows that choosing the appropriate basic model can improve the performance of the ensemble learning model.
Genetic Algorithm was combined with ensemble model for prediction of peak consumption of a building in Hong Kong the next day[11], result shows that the accuracy of the proposed model is higher than other eight individual models. Time series analysis is used to forecast the power demand [12], and the experimental results were used in the research of power consumption and electricity price prediction. Hybrid models of ARIMA and machine learning methods are proposed, it used to forecast energy demand and consumption [13-14]. Hybrid models of EMD and ARIMA[15], EMD and SVR[16] were used to long-term streamflow forecasting and electricity load forecasting. Experiment shows that the accuracy of ensemble learning and hybrid models are higher than single models.

The above-mentioned building energy prediction methods use machine learning methods, time series models, ensemble learning models and hybrid models for the prediction. Due to its non-stationarity and uncertainty, its prediction is challenging, the prediction accuracy of existing methods need to be further improved. In this study, the EA-XGBoost (EMD-ARIMA-XGBoost) model is proposed for building energy consumption forecasting, which combines EMD, ARIMA and XGBoost. It uses EMD decompose data into multiple IMFs, then ARIMA is applied to get predicted value. Finally, XGBoost after Grid-Search is used to process non-linear relations in the data.

2. Related methods

2.1. EMD
Empirical Mode Decomposition(EMD)[17] is suitable for the processing of non-linear and non-stationary time series. It usually used with other methods, such as EMD combining with BP and GA forecast wind speed[18], based on ensemble deep learning for load demand forecasting[19]. The steps are as follows:

Step1: Find all maximum points in the time series, calculate the cubic spline interpolation function, and fit all the maximum points to get $s_{\text{max}}(t)$.

Step2: Recognize the minimum points in the time series, use the above function to fit the minimum points to get $s_{\text{min}}(t)$, and calculate the mean of the max and min envelopes $m_{\text{t}}(t)$,

$$m_{\text{t}}(t) = s_{\text{max}}(t) + s_{\text{min}}(t) \over 2.$$ 

Step3: Subtract $m_{\text{t}}(t)$ from $t(t)$ to get a new sequence $c_{\text{t}}(t)$. The $c_{\text{t}}(t)$ becomes first IMF of this time series, which contains the shortest periodic component in the original sequence. Judging whether $h_{\text{t}}(t) = x(t) - m_{\text{t}}(t)$ satisfies the end condition of IMF. If not satisfied, consider $c_{\text{t}}(t)$ as a new signal $x(t)$ that needs to be decomposed again. Repeat step1 to obtain the remaining eigenmode function components.

2.2. ARIMA
Autoregressive Integrated Moving Average model (ARIMA), which predicts future data based on historical data. It can be used in time series for demand forecasting of conventional electrical load[20].

It can be expressed by equal(1):

$$x_{t} = \phi_{0} + \phi_{1}x_{t-1} + \phi_{2}x_{t-2} + \ldots + \phi_{p}x_{t-p} + \epsilon_{t} - \theta_{1}\epsilon_{t-1} - \theta_{2}\epsilon_{t-2} - \ldots - \theta_{q}\epsilon_{t-q} \quad (1)$$

Where $\phi(i = 0,1,\ldots,p)$ and $\theta(j = 1,2,\ldots,q)$ are model parameters. Assumes that the random error $\epsilon$, follows a normal distribution, and $p$ and $q$ are the orders of ARIMA.

The establish of ARIMA includes the following steps:

Step 1: If the time series is stationary, go Step 2. Otherwise, perform a d-order ordinary difference on the time series to get a stationary sequence.

Step 2: Determine the model parameters: $p$ and $q$, according to the ACF and PACF of the stationary sequence.
Step3: Parameter candidate values of this model determined according to Step1 and Step2. Calculate the AIC value of different parameter combinations separately, and select the parameter which value corresponding to the smallest AIC.

Step4: Validity test. The validity of the model uses the autocorrelation coefficient to test the residuals. If the residuals have no autocorrelation, go to Step 5. Otherwise, a new model is established with other candidate parameters until the model passes the validity test.

Step5: Forecast the future data. The fitted ARIMA model is used to predict future data.

2.3. XGBoost
XGBoost was proposed by Chen in 2016 [21]. It bases on the Classification and Regression Tree (CART), redefines the partition attributes, and uses the minimization of the loss function to determine the partition attributes. It used to forecast electricity load [22], oil price forecasting [23], forecast the accident [24], novel image classification [25] combining with CNN. Its accuracy higher than other models, and its training speed also less than other models. XGBoost minimize the loss function to determine most suitable feature, the loss function is defined by equal(2):

\[ \text{Obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \]  

(2)

Where \( L(y_i, \hat{y}_i) \) means the training error of the i-th sample, \( T \) is the number of leaf nodes, \( w \) is the score of leaf nodes and \( \gamma, \lambda \) is the coefficient.

When the k-th tree is added, the previous tree has been trained. Its training error and regular term of the previous tree become constant. The loss function is written as equal(3).

\[ \text{Obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(r-1)} + f_i(x_i)) + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 + C \]  

(3)

Assuming the loss function is Mean Square Error, and the second-order Taylor expansion of the equal(3),it can be expressed by equal(4).

\[ \text{Obj}'(\theta) \approx \sum_{i=1}^{n} g_i f_i(x_i) + \frac{1}{2} h_i f_i(x_i)^2 + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \]  

(4)

Where \( g_i = \partial_{y_i} L(y_i, \hat{y}_i^{(r-1)}), h_i = \frac{1}{2} \partial^2_{y_i} L(y_i, \hat{y}_i^{(r-1)}). \)

It can be expressed by equal(5):

\[ \text{Obj}^*(\theta) = \sum_{j=1}^{T} [(\sum_{i=1}^{n} g_i) w_j + \frac{1}{2} h_i + \lambda w_j^2] + \gamma T \]  

(5)

Where, \( w_j \) is the j-th score of this leaf node. Calculate the partial derivative about \( w_j \) for equal(5),Which can get the optimal leaf node score \( w_j^* \) and minimum loss \( \text{Obj}^* \).

3. EA-XGBoost hybrid model
ARIMA is usually used in analysis of time series, due to its accuracy and mathematical soundness, which was used for electricity price forecasting [26], wind speed and power forecasting [27] and other filed. But it cannot extract information from random terms in time series data[28]. XGBoost has obvious advantages in accuracy and training speed, but it has certain defects in learning portion features. Therefore, EA-XGBoost model is proposed to combine EMD, ARIMA and XGBoost. First, use EMD to decompose data into different IMFs, and it can be decomposed to multiple IMFs and residual. It can be expressed by equal(6):

\[ X(t) = \sum_{i=1}^{n} h_i(t) + r(t) \]  

(6)
Where $h_i(t)$ means the i-th IMF, $r(t)$ is the residual of components.

Afterwards, ARIMA model is applied for each IMF, then sum the predict result of each IMF. It can be expressed by equal(7):

$$\hat{y} = \sum_{i=1}^{n} f_i(x_1, x_2 \ldots x_{t-1}, \mathcal{E}_1, \mathcal{E}_2 \ldots \mathcal{E}_{t-q})$$

(7)

Where $x_i(1,2\ldots t-1)$ means the observed data, $\mathcal{E}_j = x_j - \hat{x}_j$ means the residual of observed data and forecast data, $p$ and $q$ are the orders of ARIMA.

At last, calculate the residual $e_i$, which can be expressed by equal (8) and use it as an input feature of XGBoost.

$$e_i = y_i - \hat{y}_i$$

(8)

$$\hat{y}_i = G(x_1, x_2 \ldots x_{t-1}, u_1, u_2 \ldots, e_j)$$

(9)

Where $x_i(i=1,2\ldots t-1)$ is the value of previous data, $u_1, u_2$ represent the energy-related factors, $e_i$ can be calculated by equal(8).

The steps of EA-XGBoost are as follows, which shows in Figure 1:

1. Remove the missing part of the dataset and divide it into training set and test set.
2. Use EMD to decompose the training data into several IMFs, assuming IMF1, IMF2 ... IMFn.
3. ARIMA model is applied for each IMF, and d-order difference is made to the model to make data stationary. Determine the model parameters: $p$ and $q$, according to the ACF and PACF of the stationary sequence.
4. Use the fitted ARIMA to predict the data of each IMFs, and get predict result from EMD-ARIMA, $\hat{y}_1, \hat{y}_2 \ldots \hat{y}_n$.
5. Sum the $\hat{y}_i(1,2\ldots n)$ that come from(4) to get the $\hat{y}$, and subtract $\hat{y}$ from $y$ to get the residual $e_i$.
6. Build XGBoost model, which Grid-Search is used to find optimal hyperparameters. Using $e_i$ and other energy-related factors, the training data is used to train the model.
7. Use the test data to test the accuracy of the EA-XGBoost model, and calculate MAE, RMSE, and $R^2$ as evaluation metrics for model performance.

4. Evaluation metrics

MSE, RMSE and $R^2$[29-30] are adopted in this paper. They are defined by equations (10-13), given in Table 1.

| Evaluation metrics | Formula |
|--------------------|---------|
| $MSE$              | $\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$ |
| $RMSE$             | $\sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$ |
| $MAE$              | $\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$ |
| $R^2$              | $1 - \frac{\sum_{i} (\hat{y}_i - \bar{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$ |

Table 1. Evaluation metrics
\( n \) is the number of samples, \( \hat{y}_i \) is the predicted value of the \( i \)-th sample, \( y_i \) is the observed value of the \( i \)-th sample, \( \bar{y} \) is the mean of all samples.

![Diagram](image_url)

**Figure 1.** This is the steps of proposed methods

5. **Experimental results**
   In this study, the dataset is provided by US National Renewable Energy Laboratory, it concludes the hourly energy consumption data of a building in a year. In the following contrast experiments, the above dataset was split into 10 months training set and 2 months for test, use the former set to train EMD-ARIMA, XGBoost and EA-XGBoost three models, and predict final 2 months' building energy consumption. The three models are compared with the evaluation metrics mentioned above.

5.1 **Experiment of EMD-ARIMA**
   First, we applied EMD in the training set to decompose it into 6 IMFs as shown in the Figure 2. Then for the 6 IMFs, sum the ARIMA prediction result as the final result. Take the first IMF for example, and make it stationary. According to the ACF and PACF, we can determine the suitable value of \((3,1,3)\) for this model. As it shown in Figure3(1), after determining the parameters, use fitted ARIMA to predict. Finally, sum the prediction of each IMF to get the result, the prediction result of the model are shown in Figure 4(a). Evaluation metrics are shown in Table 2. Because EMD-ARIMA model only extracts the time information from the data and ignores the energy-related factors such as the dry and wet bulb temperature, the accuracy for prediction can be improved.
Figure 2. EMD components of building energy consumption dataset

Figure 3. Autocorrelation (ACF) and Partial Autocorrelation (PACF) plots for dataset
5.2 Experiment of XGBoost
In this part, firstly, we used the Grid-Search to optimize the hyperparameters of XGBoost model. It was determined that the value of eta for 0.1, and the subsample for 0.8, the colsample bytree for 0.85 and max_depth for 10. Then the XGBoost after Grid-Search model was applied for the dataset, the prediction results of the model are shown in Figure 4(b), and the evaluation metrics are shown in Table 1. XGBoost considered the important energy-related characteristics such as dry and wet bulb temperature, so the performance of the model is improved compared with EMD-ARIMA model. However, XGBoost does not fully extract the time characteristics of building energy consumption, so there still can improve the accuracy for its prediction.

5.3 Experiment of EA-XGBoost
To solve the above problem, the EA-XGBoost model was proposed for the building energy consumption forecasting. The new model uses EMD-ARIMA to fit the data and calculates the residual value as a new feature of XGBoost. It used the XGBoost model after Grid-Search to predict on the test datasets. The prediction results of EA-XGBoost are shown in Figure 4(c), and the evaluation metrics are shown in Table 1. The experimental results show that EA-XGBoost model fully considers the time factor of data and other energy-related factors. The prediction accuracy has been greatly improved, compared with other two models, the EA-XGBoost shows obvious advantages in the evaluation metrics.

5.4 Experimental Results
As shown in Table 1, compared with EMD-ARIMA model, the MAE of EA-XGBoost model decreased by 41.657, the RMSE decreased by 41.544, and the $R^2$ increased by 0.18. Compared to XGBoost, MAE decreased by 11.728, RMSE decreased by 11.584, and $R^2$ increased by 0.039. From Figure 4, the EA-XGBoost model has higher accuracy than EMD-ARIMA and XGBoost.

In summary, the prediction performance of EA-XGBoost model in building energy consumption prediction is better than EMD-ARIMA and XGBoost, which verifies the feasibility and practicability of the proposed EA-XGBoost model.
Figure 4. EMD-ARIMA, XGBoost and EA-XGBoost models’ prediction results on the building energy consumption dataset.

Table 2. Performance result of three models

| Model        | MAE  | RMSE  | R²    |
|--------------|------|-------|-------|
| EMD-ARIMA    | 88.477 | 88.558 | 0.749 |
| XGBoost      | 58.548 | 58.598 | 0.890 |
| EA-XGBoost   | 46.820 | 47.014 | 0.929 |
6. Conclusion
EA-XGBoost is proposed in this study, which combines EMD, ARIMA and XGBoost. It uses EMD to decompose data into multiple IMFs, and applies ARIMA for each IMF. Sum the predict result from each IMF, and calculate the residual, which is taken as an input feature of XGBoost. The proposed model is compared with EMD-ARIMA and XGBoost model, and the evaluation metrics MAE, RMSE and $R^2$ are adopted to measure above models on the dataset of building energy consumption provided by the US National Energy Laboratory. The contrast experimental result shows that the superiority of the proposed model. Compared with the EMD-ARIMA and XGBoost, its MAE and RMSE has a decrease of 41.657, 41.544 and 11.728, 11.584, and $R^2$ increase 18% and 3.9%. The proposed hybrid model can provide energy-saving decision-making and energy management support for government agencies, enterprises and institutions. It can be used in other field, such as electricity price forecasting, streamflow forecasting and finance-related field.

There are many factors on building energy consumption. In this study, the feature used in the EA-XGBoost model are limited to other related researches, which may limit the prediction accuracy of the model. As a future work, data mining methods can be used to find other energy-related factors. Besides, in future, multiple-step forecast will be considered.

References
[1] L. Pérez-Lombard, J. Ortiz, and C. Pout, “A review on buildings energy consumption information,” Energy and Buildings, vol. 40, no. 3, pp. 394–398, Jan. 2008.
[2] J. Zhan, W. Liu, F. Wu, Z. Li, and C. Wang, “Life cycle energy consumption and greenhouse gas emissions of urban residential buildings in Guangzhou city,” Journal of Cleaner Production, vol. 194, pp. 318–326, Sep. 2018.
[3] K. Amasyali and N. M. El-Gohary, “A review of data-driven building energy consumption prediction studies,” Renewable and Sustainable Energy Reviews, vol. 81, pp. 1192–1205, Jan. 2018.
[4] A. Alshibani and O. S. Alshamrani, “ANN/BIM-based model for predicting the energy cost of residential buildings in Saudi Arabia,” Journal of Taibah University for Science, vol. 11, no. 6, pp. 1317–1329, Nov. 2017.
[5] D. Madronal, R. Lazcano, H. Fabelo, S. Ortega, R. Salvador, G. M. Callico, E. Juarez, and C. Sanz, “Energy consumption characterization of a Massively Parallel Processor Array (MPPA) platform running a hyperspectral SVM classifier,” 2017 Conference on Design and Architectures for Signal and Image Processing (DASIP), Sep. 2017.
[6] E. Vinagre, T. Pinto, S. Ramos, Z. Vale, and J. M. Corchado, “Electrical Energy Consumption Forecast Using Support Vector Machines,” 2016 27th International Workshop on Database and Expert Systems Applications (DEXA), Sep. 2016.
[7] A. S. Ahmad, M. Y. Hassan, M. P. Abdullah, H. A. Rahman, F. Hussin, H. Abdullah, and R. Saidur, “A review on applications of ANN and SVM for building electrical energy consumption forecasting,” Renewable and Sustainable Energy Reviews, vol. 33, pp. 102–109, May 2014.
[8] M. Torabi, S. Hashemi, M. R. Saybani, S. Shamshirband, and A. Mosavi, “A Hybrid clustering and classification technique for forecasting short-term energy consumption,” Environmental Progress & Sustainable Energy, vol. 38, no. 1, pp. 66–76, Jun. 2018.
[9] M. W. Ahmad, M. Moursheed, and Y. Rezgui, “Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption,” Energy and Buildings, vol. 147, pp. 77–89, Jul. 2017.
[10] Z. Wang and R. S. Srinivasan, “A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models,” Renewable and Sustainable Energy Reviews, vol. 75, pp. 796–808, Aug. 2017.
[11] C. Fan, F. Xiao, and S. Wang, “Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques,” Applied Energy, vol. 127, pp. 1–10, Aug. 2014.
[12] S. S. Pappas, L. Ekonomou, P. Karampelas, D. C. Karamousantas, S. K. Katsikas, G. E. Chatzarakis, and P. D. Skafidas, “Electricity demand load forecasting of the Hellenic power system using an ARMA model,” Electric Power Systems Research, vol. 80, no. 3, pp. 256–264, Mar. 2010.

[13] G. P. Zhang, “Time series forecasting using a hybrid ARIMA and neural network model,” Neurocomputing, vol. 50, pp. 159–175, Jan. 2003.

[14] H. Cui, P. Wei, Y. Mu, and X. Peng, “SARIMA-Orthogonal Polynomial Curve Fitting Model for Medium-Term Load Forecasting,” Discrete Dynamics in Nature and Society, vol. 2016, pp. 1–9, 2016.

[15] Z.-Y. Wang, J. Qiu, and F.-F. Li, “Hybrid Models Combining EMD/EEMD and ARIMA for Long-Term Streamflow Forecasting,” Water, vol. 10, no. 7, p. 853, Jun. 2018.

[16] Y. Yaslan and B. Bican, “Empirical mode decomposition based denoising method with support vector regression for time series prediction: A case study for electricity load forecasting,” Measurement, vol. 103, pp. 52–61, Jun. 2017.

[17] N. E. Huang, “An Adaptive Data Analysis Method for Nonlinear and Nonstationary Time Series: The Empirical Mode Decomposition and Hilbert Spectral Analysis,” Applied and Numerical Harmonic Analysis, pp. 363–376.

[18] S. Wang, N. Zhang, L. Wu, and Y. Wang, “Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method,” Renewable Energy, vol. 94, pp. 629–636, Aug. 2016.

[19] X. Qiu, Y. Ren, P. N. Suganthan, and G. A. J. Amaratunga, “Empirical Mode Decomposition based ensemble deep learning for load demand time series forecasting,” Applied Soft Computing, vol. 54, pp. 246–255, May 2017.

[20] M. H. Amini, A. Kargarian, and O. Karabasoglu, “ARIMA-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation,” Electric Power Systems Research, vol. 140, pp. 378–390, Nov. 2016.

[21] T. Chen and C. Guestrin, “XGBoost,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, 2016.

[22] H. Zheng, J. Yuan, and L. Chen, “Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation,” Energies, vol. 10, no. 8, p. 1168, Aug. 2017.

[23] M. Gumus and M. S. Kiran, “Crude oil price forecasting using XGBoost,” 2017 International Conference on Computer Science and Engineering (UBMK), Oct. 2017.

[24] X. Shi, Q. Li, Y. Qi, T. Huang, and J. Li, “An accident prediction approach based on XGBoost,” 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nov. 2017.

[25] X. Ren, H. Guo, S. Li, S. Wang, and J. Li, “A Novel Image Classification Method with CNN-XGBoost Model,” Lecture Notes in Computer Science, pp. 378–390, 2017.

[26] J. Contreras, R. Espinola, F. J. Nogales, and A. J. Conejo, “ARIMA models to predict next-day electricity prices,” IEEE Transactions on Power Systems, vol. 18, no. 3, pp. 1014–1020, Aug. 2003.

[27] R. G. Kavasseri and K. Seetharaman, “Day-ahead wind speed forecasting using f-ARIMA models,” Renewable Energy, vol. 34, no. 5, pp. 1388–1393, May 2009.

[28] Aasim, S. N. Singh, and A. Mohapatra, “Repeated wavelet transform based ARIMA model for very short-term wind speed forecasting,” Renewable Energy, vol. 136, pp. 758–768, Jun. 2019.

[29] Manikanta C and Mamatha Jadav V, “Evaluation of modified PLS regression method to fill the missing values in training dataset,” 2015 International Conference on Smart Sensors and Systems (IC-SSS), Dec. 2015.

[30] Bin Sun, Wei Cheng, P. Goswami, and Guohua Bai, “Flow-aware WPT k-nearest neighbours regression for short-term traffic prediction,” 2017 IEEE Symposium on Computers and Communications (ISCC), Jul. 2017.