2021 IEEE 9th International Conference on Smart Energy Grid Engineering (SEGE) | 978-1-6654-4094-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/SEGE52446.2021.9534987

Sampling Strategy Analysis of Machine Learning Models for Energy Consumption Prediction

Zeqing Wu
Thermal and Power Engineering Department
Tianjin University
Tianjin, China
wzqnb@tju.edu.cn

Weishen Chu*
Walker Department of Mechanical Engineering
The University of Texas at Austin
Austin, United States
wschu@utexas.edu

***Abstract*—With the development of the Internet of things (IoT), energy consumption of smart buildings has been widely concerned. The prediction of building energy consumption is of great significance for energy conservation and environmental protection as well as the construction of smart city. With the development of artificial intelligence, machine learning technology has been introduced to energy consumption prediction. In this study, multiple learning algorithms including Support Vector Regression (SVR), Artificial Neural Network (ANN), Random Forest (RF) are developed to perform energy consumption prediction. The most appropriate machine learning algorithm for energy consumption prediction has been investigated and found to be the random forest algorithm. Based on the developed machine learning models, studies on the sampling strategy for energy consumption prediction have been conducted. It is found that the variance of data has a significant effect on the prediction accuracy, and a better prediction result can be achieved by increasing the sampling density over the data with high variance. This result can be used to optimize the machine learning algorithm for building energy consumption prediction and improve the computational efficiency.**

Keywords—energy consumption, machine learning, random forest, sampling strategy

I. Introduction

Energy consumption of smart buildings has been drawing attention from the Internet of things (IoT) field [1]. It is widely known that the amount of building energy consumption is huge [2], [3]. For example, in America and the European Union, the energy demand for residential and commercial buildings accounts for about 40% of the total energy demand. In China, the energy demand for residential and commercial buildings is as high as 30%. About 63% of energy consumption is for the building's internal heating and refrigeration [4]. The huge energy consumption can also bring the ascension of emissions of greenhouse gases such as carbon dioxide. According to research, in the United States, building energy consumption accounts for 40% of total CO_2 emissions [5]. Therefore, it is of great importance to develop a predictive model to predict the energy consumption of a smart building. Based on the predictive model, building administrators are enabled to carry on the demand side management (DSM) and the demand side response (DSR) [6]. In the field of environment engineering, effective energy consumption prediction models enable researchers to better control the overall energy consumption of buildings, so as to achieve the purpose of reducing greenhouse gas emissions.

Researchers have carried out a lot of research on smart building energy consumption prediction [7], [8]. However, the research on building energy consumption prediction is still in its infancy, and there are some deficiencies. For example, in the process of prediction, a proper methodology is needed to identify the best prediction method for effective prediction with respect to various situations, including different time periods and different regions of buildings. Also, sampling strategies for machine learning models on energy consumption prediction have not been deeply studied.

To resolve the problems discussed, this paper developed four machine learning models to predict the energy consumption of a smart building and evaluate the performance of each developed model, based on the data recording of an energy-saving house [6]. The recorded data included the temperature, humidity, weather stations and the electricity consumption of electric lights. The relationship and dependence between the variables and their effects on the housing energy consumption were investigated, and the effects of sampling density on the prediction performance were also studied. Based on those findings, a machine learning sampling strategy for energy consumption prediction was proposed and verified.

II. Machine Learning Algorithms and Prediction Error Metrics

In this study, three machine learning algorithms and one polynomial regression algorithm have developed to conduct energy consumption prediction. The principles of those algorithms used in this study are introduced in this section.

A. Support Vector Machines

Support vector machine algorithm [9] is suitable for small and medium data sets. This algorithm can be used for both classification and regression prediction analysis. The regression problem can be described as follows.

Given a data set $D = \{(x_1, y_1), (x_2, x_2), \ldots (x_n, y_n)\}$ for perform regression prediction, the function $f(x) = w^T \cdot x + b$ needs to be fitted, so that the value of $f(x)$ is as close to the value of y as possible.

The principle of the support vector regression algorithm can be expressed by (1)

$$f(x) = \sum_{i=1}^{n} (\hat{\alpha}_i - \alpha_i) \Phi(x_i)^T \Phi(x_i) + b \qquad (1)$$

In equation (1), $\Phi(x)$ represents the kernel function. There are several types of kernel functions that are often used, namely polynomial kernel function and Gaussian radial basis function (RBF) [9]. The kernel function affects the accuracy of the support vector machine for regression prediction. The support vector machine algorithm is very sensitive to the scaling of features and effective for small-scale data sets.

B. *Artificial Neural Networks*

Artificial neural network (ANN) algorithm is an algorithm that uses artificial structure to simulate the structure and function of human brain [10]. The complex network structure of this algorithm is based on a large number of simple neurons [11]. Therefore, this algorithm is highly nonlinear, which enables ANN model to conduct complex logic operations and simulate nonlinear relations. Fig. 1 shows an illustration of an ANN structure.

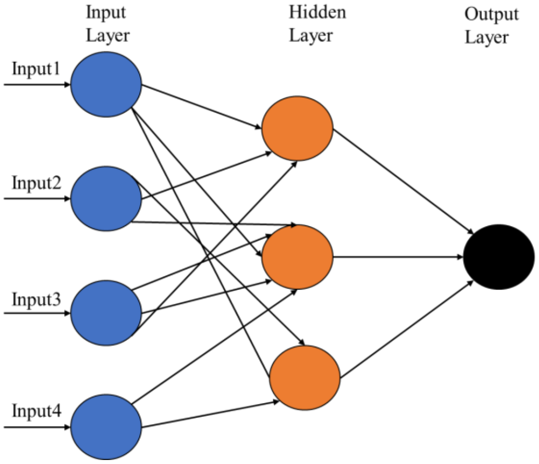


Fig. 1. Illustration of artificial neural network.

The neural network consists of an input layer, hidden layers and an output layer. The hidden layer is composed of various neurons, which enable the neural network to learn the complex relationship between inputs and outputs form training data.

For a single neuron model, Fig. 2 shows its structure and principle.

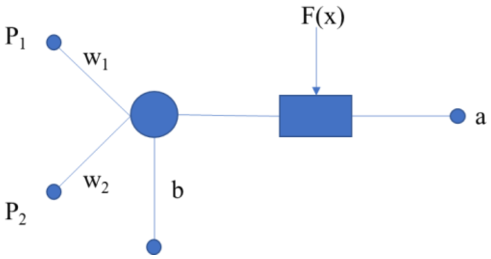


Fig. 2. Structure of neuron from ANN model.

The signal transmission process in the ANN model can be express as (2)

$$A = f(W \cdot P + b) = f(\textstyle\sum_{i=1}^{2} w_i \cdot p_i + b) \qquad (2)$$

where *f(x)* is the activation function to calculate the output signal from a neuron, *P* is the signal received by the neuron. A weight *W* and bias *b* are introduced to conduct the model training and optimization.

Therefore, the activation function has a great influence on the performance of neural network algorithm. The activation function used in this study is ReLU (rectified linear units) function, which is widely used in supervised learning, as expressed in (3).

$$Y = \begin{cases} W \cdot P + b & W \cdot P + b \geq 0 \\ 0 & W \cdot P + b \geq 0 \end{cases} \qquad (3)$$

The RLU function has the characteristic that the input and output of the neuron satisfy the linear relationship in a certain interval as shown in Fig. 3, which simulates the saturation characteristics of the actual system.

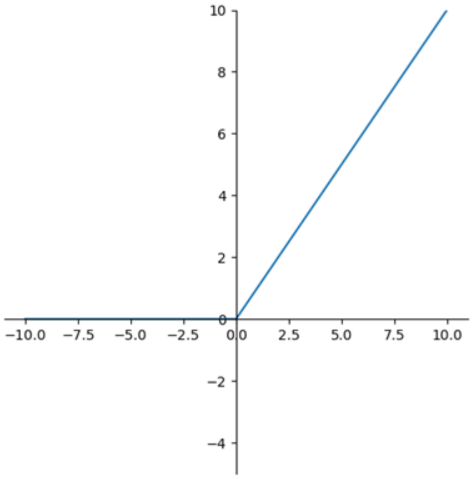


Fig. 3. Illutrstion of ReLU function.

C. *Random Forests*

The Random forest algorithm belongs to Bagging algorithm [12]. This integrated learning can be understood as a packaging of multiple weak models to form a strong model. In the random forest algorithm, the weak model employs the decision tree algorithm and trains multiple decision trees to form a forest. The CART algorithm in Sklearn package is applied to generate binary trees. The random forest algorithm can be represented by the (4).

$$G(x_i, v_{i,j}) = \frac{n_{left}}{N_s} H(X_{left}) + \frac{n_{right}}{N_s} H(X_{right}) \qquad (4)$$

where x_i is a categorical variable, and $v_{i,j}$ is the split value of the categorical variable. n_{left} is the number of training samples of the left child node after segmentation. n_{right} is the number of training samples of the right child node after segmentation. N_s is the number of all training samples of the current node. X_{left} and X_{right} are the training sample sets of the left and right child nodes respectively. $H(X)$ is a function to measure the impurity of the node.

D. Prediction Error Metrics

This article uses the following parameters to evaluate the accuracy of prediction.

1. Mean Square Error (MSE)

The mean absolute error is defined as follows

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(yf^{(i)} - y^{i})^2 \qquad (5)$$

where m is the number of data points, $yf^{(i)}$ is the i-th predicted value, and y^i is the i-th actual value. The MSE tends to assign more weights to larger errors and therefore is more appropriate when larger errors are especially undesirable.

2. Mean Absolute Error (MAE)

Mean absolute error is the average of the absolute value of the difference between the predicted value and the actual value, which can be expressed in (6).

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|yf^{(i)} - y^{i}\right| \qquad (6)$$

where m is the number of data points, $yf^{(i)}$ is the i-th predicted value, and y^i is the i-th actual value. MAE and MSE are not monotonic with each other, and it is possible that one increases at the decrease of the other.

III. Content and Results of the Experiment

A. Data Source and Pre-processing

The datasets used in this study are from UCI's public database [6]. The data set recorded the energy consumption of a residential building in Stambruges, Belgium. The residence was completed in December 2015, with a total construction area of 280 square meters, of which the heating area is 220 square meters. Inside the house, sensors were used to record relevant physical quantities, including electrical consumption, lighting consumption, temperature and humidity in each room. Meteorological conditions were recorded by the airport meteorological station. The time interval for recording data was ten minutes. The recording time span was from 17:00 on January 11, 2016 to 18:00 on May 27, 2016. The major focus of the experiment is to use the relevant parameters to predict the consumption of electrical appliances in the house. Fig. 4 shows the distribution of electrical consumption over time.

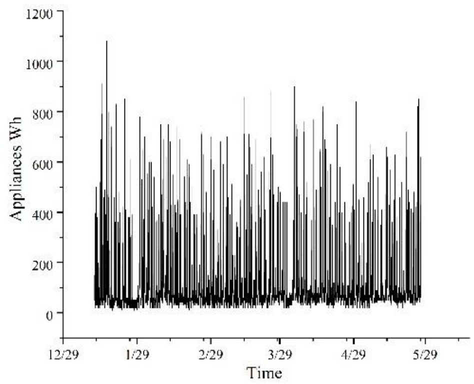


Fig. 4. Distribution of electrical consumption over time.

Before developing machine learning models, we analyzed the correlation coefficients of each independent variable to the dependent variable and screened them to realize an effective training. Fig. 5 shows the distribution of each correlation coefficient.

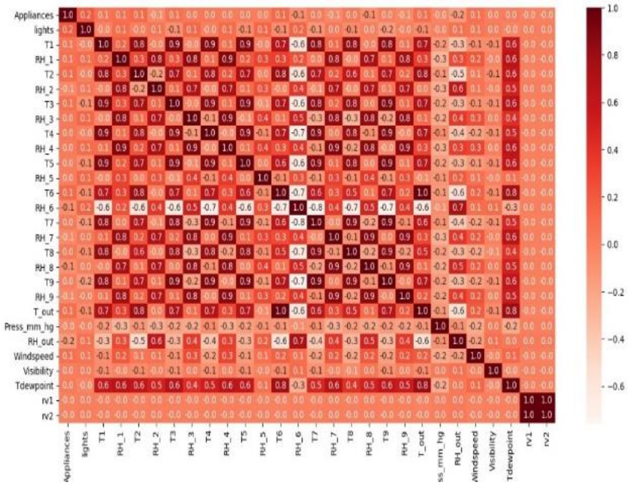


Fig. 5. Heatmap of correlation coefficient matrix.

By defining the threshold of correlation coefficient to be 1 , we selected *lights, T1, RH_ 1, T2, RH_ 2, T3, T6, RH_ 6, RH_ 7, RH_ 8, RH_ 9, T_out, RH_ out* and *Windspeed* from the database as independent variables. The total number of independent variables is 14. By choosing *Appliances* as the dependent variable to train machine learning algorithm, we developed four models to predict the energy consumption, including Linear Regression, Support Vector Regression, Artificial Neural Network and Random Forest. 95% of the recorded data is used to train the algorithm and 5% to verify the accuracy of the models.

B. Performance of Machine Learning Models

The prediction results of the four models are shown in Fig. 6, where predicted and validation appliances over time are both plotted to demonstrate the prediction performance.

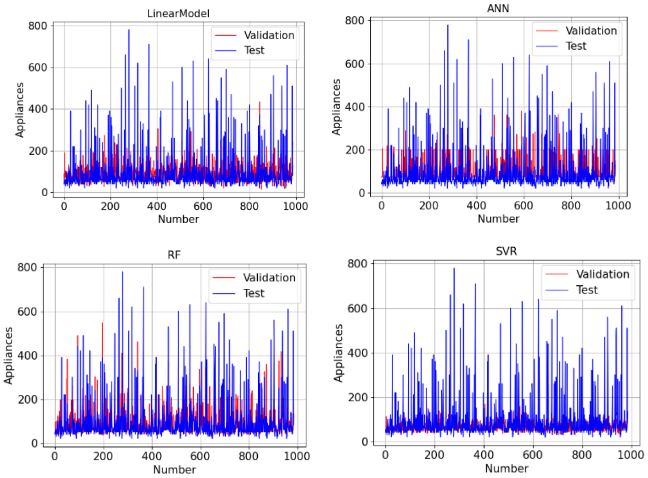


Fig. 6. Distribution of predicted results from the different machine learning model.

The results from Fig. 6 provide an intuitive comparison between the predicted electrical energy consumption and the actual electrical energy consumption, where the results from the Random Forest model are found to have the most overlap with the validation data. In order to quantitively evaluate the performance of each algorithm, the mean absolute error (MAE) and mean square error (MSE) of each model has been calculated and plotted in Fig. 7.

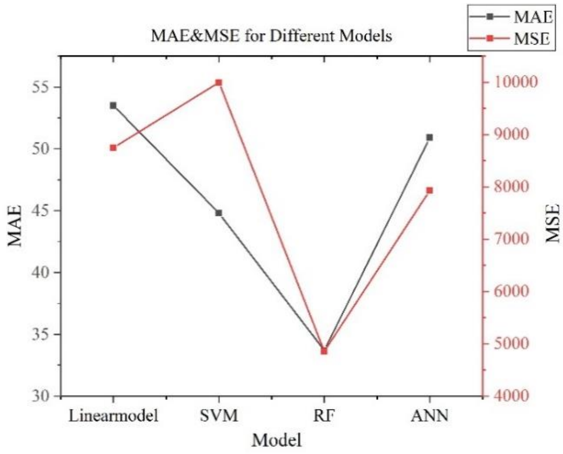


Fig. 7. MAE and MSE for developed models.

By comparing the MAE and MSE of the four models, it can be found that the random forest algorithm performs better than the other algorithms. The MAE and MSE are about 50% lower than other algorithms. Therefore, we chose random forest algorithm as the algorithm for further experiments.

C. Sampling Stratgery to Improve Predictive Performance

The experiments discussed are based on the analysis over the entire data set. We have concluded that the random forest algorithm is most appropriate model for energy consumption prediction. We further found that the fluctuation of electrical power consumption varies at different time of the day. Taking the electric power consumption data of January 14, 2016 as an example, we divided the time period into three groups: 2:00-4:00, 12:00-14:00, and 20:00-22:00 to represent early morning, noon and evening. On this basis, the relationship between electrical power consumption and time period can be shown in Fig. 8.

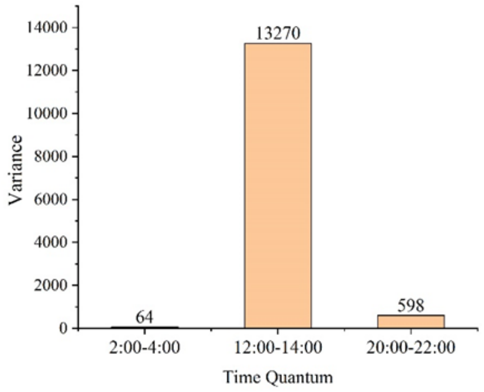


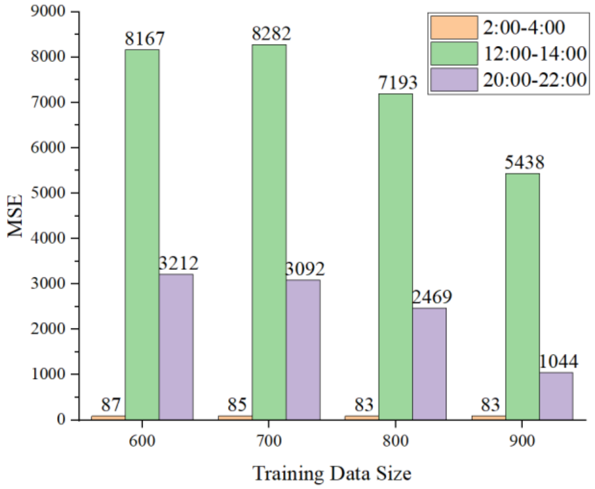
Fig. 8. Varence of difference time period.

Based on Fig. 8, we found that at the time period of 2:00-4:00, the fluctuation of electrical power consumption is very small with the variance of 64, while at the time period of 12:00-14:00, the variance is very large, reaching 13270, which indicates that the fluctuation of electrical power consumption is very large. The fluctuation of electrical power consumption between 20:00-22:00 in the evening falls in between of the other two periods.

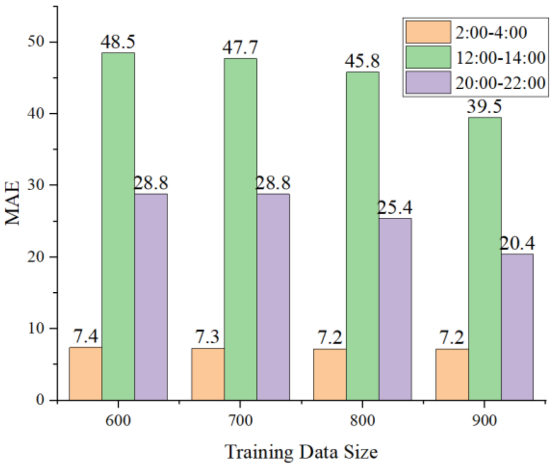
The different variances for those three time periods guided us to consider assigning different amount of training data for each period, because features with higher variance are usually harder to be captured by the model compared to those with lower variance. A sampling strategy can be drawn accordingly by collecting more data over the noon period and reducing the train data set over the morning period to improve the accuracy and efficiency.

To verify this variance-based sampling strategy for energy prediction, we proceeded to use the random forest algorithm to fit the data of different time periods with different amount of training data. The amount of training data was set to be 600, 700, 800, and 900 respectively for each time period. MAE and MSE were employed to evaluate the prediction performance.

As shown in Fig. 9 (a), given the same training data size, the MAE of the model at the three different time periods are not the same, where it minimizes at the period of 2:00-4:00 and maximizes at the period of 12:00-14:00. Based on the results in Fig. 9 (a), it can be found that the MAE of 12:00-14:00 case can be as large as seven times higher than that of 2:00-4:00 case, given the same amount of training data over each time period. This finding has verified the effect of sample variance on the performance. We also found that at the time period of 2:00-4:00, the value of MAE does not reduce as significantly as the increase of the amount of test data, and it keeps at a relatively stable value. The MAE values of other time periods, however, show an obvious downward trend with the increase of the amount of training data. This finding verifies that it is more efficient to boost the amount of training data over the time period with large variance instead of small variance. The similar trend has been found in MSE analysis in Fig. 9 (b), which further verified the proposed sampling strategy.



(a)



(b)

Fig. 9. MAE (a) and MSE (b) for models with different training data size.

Based on the analysis above, we have verified the methodology to collect training data based on the variance of the sample that weights more on the case with higher fluctuation. This achieves reasonable and effective training data preparation so as to save computer computing resources and make accurate prediction of electrical energy consumption.

IV. Conclusion and Future Work

In this paper, we have developed four different types of machine learning model (LR, SVR, ANN, RF) to predict the energy consumption of electrical appliances in an energy-saving building. By comparing the MAE and MSE of the developed models, we find that the Random Forests model has the best prediction performance with the highest accuracy.

Based on the pre-processing on the data set, we find that the fluctuations of the data at different time periods are inconsistent. A variance-based sampling strategy is then developed to improve the prediction accuracy and computational efficiency. The developed sampling strategy has been verified by studying the effect of sample variance on MSE and MAE. This sampling strategy together with the identified machine learning model can effectively save computer computing resources and improve the efficiency of the energy consumption prediction system. Future work can include further investigation on gradient-based sampling strategies. The machine learning model can also be extended to deep learning field to handle high-dimensional input scenario.

References

[1] Jia, M., Komeily, A., Wang, Y., & Srinivasan, R. S. (2019). Adopting Internet of Things for the development of smart buildings: A review of enabling technologies and applications. Automation in Construction, 101, 111-126.

[2] Molina-Solana, M., Ros, M., Ruiz, M. D., Gómez-Romero, J., & Martín-Bautista, M. J. (2017). Data science for building energy management: A review. Renewable and Sustainable Energy Reviews, 70, 598-609.

[3] Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., & Pendyala, R. M. (2017). Machine learning approaches for estimating commercial building energy consumption. Applied energy, 208, 889-904.

[4] Hamilton, Ian, Oreszczyn, Tadj, Huebner, & Gesche M.etc. (2015). Explaining domestic energy consumption - the comparative contribution of building factors, socio-demographics, behaviours and attitudes. Applied energy.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[5] Zhao, Hai-xiang, and Frédéric Magoulès. "A review on the prediction of building energy consumption." Renewable and Sustainable Energy Reviews 16.6 (2012): 3586-3592.

[6] Candanedo, L. M., Feldheim, V., & Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. Energy and buildings, 140, 81-97.

[7] Qiuhong, Z., Zhan, Z., & Jiayi, W. (2020, January). Application of Regional Building Energy Consumption Prediction Model in Building Construction. In 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS) (pp. 92-94). IEEE.

[8] Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews, 81, 1192-1205.

[9] Olanrewaju, O. A. (2019, December). Predicting Industrial Sector's Energy Consumption: Application of Support Vector Machine. In 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 1597-1600). IEEE.

[10] Chu, Weishen. Studies on the effects of wiring density on chip package interaction and design optimization with machine learning. Diss. 2021.

[11] Jin, B., Cruz, L., & Gonçalves, N. (2020). Deep facial diagnosis: deep transfer learning from face recognition to facial diagnosis. IEEE Access, 8, 123649-123661.

[12] Darlis, Dila Najwa, et al. "Random Forest Approach for Energy Consumption Behavior Analysis." 2020 IEEE Symposium on Industrial Electronics & Applications (ISIEA). IEEE, 2020