Building Energy Consumption Prediction with Principal Component Analysis and Artificial Neural Network

Mengxuan Sun[1], Jinglin Zhao[2], and Heidan Shang[3]
[1]University of Essex, Colchester, UK
[2]Hong-Kong Baptist University, Kowloon Tong, Hong Kong
[3]Hydraulician at Ningxia Water Resources & Hydropower Survey Design & Research Institute, Ningxia, China
Email: sunmengxuan7@163.com, {Jlinzhao, Shangheidan}@outlook.com

***Abstract*—The implementation of the smart grid will greatly improve the efficiency of energy supply by detecting, predicting, and reacting to real-time local changes of energy uses. To this end, energy usage prediction of household buildings is critically important to facilitate the implementation of smart grid. This study used a single house as a prototype, employed different observed features, advanced data analysis approach, and artificial neural network model to predict real-time dynamics of house energy usage. Data analysis revealed that among the 26 observed features, only the top ten most important features were helpful and could maximize the neural network model performance. The resultant model has the great predictive capability on energy usage, thus provided a promising framework to improve the smart grid implementation.**

***Index Terms*—building energy use, machine learning, principal component analysis, recurrent neural network**

I. Introduction

Smart Grid (SG) is an intelligent electricity network based on an integrated, high-speed two-way communication network, through advanced sensing and measurement technology, advanced equipment technology, advanced control methods and the application of advanced decision support system technologies to achieve reliable, safe, economical, efficient, environmentally friendly and safe use of the grid is promote. SG aims to implement a network that could efficiently distribute energy directly from the power plant to locations of home or business. Therefore, understanding the dynamics of power generation and energy demand in different regions is critically important for energy delivery efficiency [1]. Residential building energy usage occupied about 20% - 30% of total electric energy demand [2], [3], thus plays an important role in smart grid implementation. For example, a study of residential buildings in the UK indicated that electricity consumption in television and other electronic appliances operating on standby increased by 10.2%. By integrating the actions of all users connected to it and utilizes advanced information, control and communication technologies, SG is able to largely save electric energy, reduce costs, and improve reliability and transparency [4].

For household building, interior and exterior factors both significantly control the overall building energy use. From a domestic aspect, two main issues are critical: type of electric appliances and the usage of them. External factors such as temperature, humidity, light and time are usually strongly related to the usage of electronic devices. For instance, the air-conditioner is a high power consuming machine. During summer time when the temperature is high, residential electrical energy will increase sharply because of using the air-conditioner. Other studies also highlighted that the end-user load exhibited important temporal feature, e.g., cooking (food preparation) dishwashers, lamps, and small appliances energy usage showed a significant evening peak [5], [6].

In order to establish a relationship between observable features and the building energy use and make predictions on future energy use, data-driven machine learning models have been widely used. Regression models, engineering methods, support vector machines [7], multiple regression, neural networks, forecasting methods [8]-[10], Hidden Markov model [11], time series analysis [12] all have been applied to predict the electricity usage or demand. However, a consistent and efficient workflow for selecting useful predictors and using them in advanced machine learning model to generate reliable energy use prediction has been achieved. In this paper, data feature analysis technique and neural networks models are combined, standardized, and trained with detailed building energy consumption measurements. Our objective is to test the efficacy of the combined modeling framework, thus provide an opportunity to facilitate and improve the implementation of the smart grid.

II. Methodology

Detailed energy use data as well as multiple interiors, exterior environmental features were considered in this study [13]. This dataset was based on a house in Stambruges, which is a low-energy house [14] completed

Manuscript received January 16, 2020; revised May 22, 2020.

doi: 10.18178/ijeee.8.2.36-39

in December 2015. Energy use (Wh) and temperature and humidity in multiple rooms were recorded every 10 minutes. For the exterior environment of the house, pressure, humidity, temperature, wind speed, visibility was obtained from a local weather station. We smoothed the date using a 24-hour filter, in order to make daily scale energy use prediction rather than 10 minutes.

To build and train an effective machine learning model, one of the greatest challenges is to provide the model useful but non-redundant features as predictors. In this case, we employed Principal Component Analysis (PCA), which transforms the original data into a set of linearly independent representations of each dimension and thus extract the main and non-redundant features of the input data. In PCA transformation, data are converted from the original coordinate system to a new coordinate system, and the selection of the new coordinate system is determined by the data itself. The first new axis selects the direction with the largest variance in the original data, and the second new axis selects the direction orthogonal to the first coordinate axis and having the second largest variance. This process is repeated until the number of repetitions equal to the number of features in the original data. It has been observed that most of the variances are included in the first few new axes. Therefore, the remaining axes can be ignored. In this study, the input feature contains twenty-six different variables. Fig. 1 showed that some features were significantly related to others (deep red was positively related and deep blue was negatively related). Thus, reduce redundant input features became necessary.

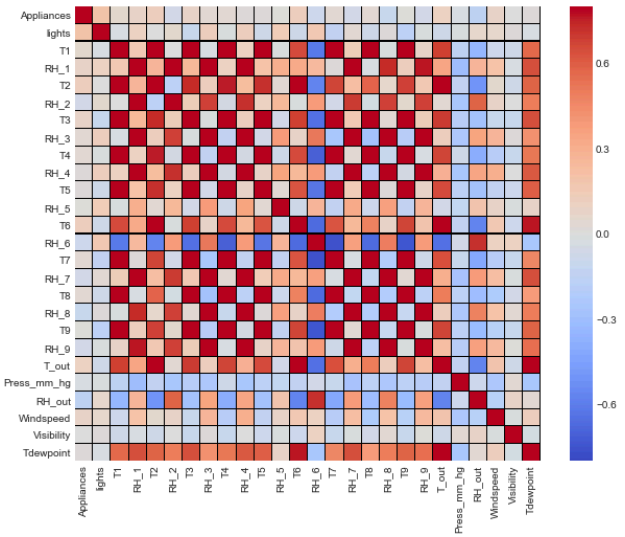

Figure 1. Pearson correlation among of input features an output target.

After feature selection, we employed Recurrent Neural Network (RNN) (architecture showed in Fig. 2), in which nodes are connected in the direction of a time series to form a directed graph, to model and predict the house energy use in the future. Unlike feedforward neural networks, recurrent neural networks can process sequences by hiding states and optional outputs, looping through the inner networks.

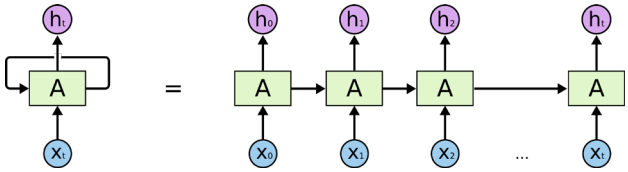

Figure 2. Structure of recurrent neural network.

However, simple recurrent neural networks cannot handle long-term time correlations because of recursion, weight exponential explosion or vanishing gradient problems. Long short-term memory LSTM, a specific RNN architecture, is more suitable for processing and predicting important events with relatively long memory [15]. Unlike traditional RNN, LSTM adds one or more memory cells to each node, including input gates, forgetting gates, and output gates, which can determine if the information is useful and connect previous information to the current task. Thus, the model for predicting building energy use with longtime series was decided to adapt LSTM recurrent neural network. Fig. 3 demonstrated the structure of one node in LSTM architecture

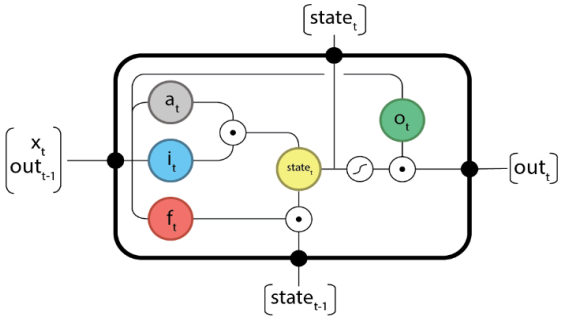

Figure 3. The inner structure of a recurrent neural network with LSTM architecture.

Input activation

$$a_t = \tanh(W_a * x_t + U_a * out_{t-1} + b_a) \quad (1)$$

Input gate

$$i_t = \sigma(W_i * x_t + U_i * out_{t-1} + b_i) \quad (2)$$

Forget gate

$$f_t = \sigma(W_f * x_t + U_f * out_{t-1} + b_f) \quad (3)$$

Output gate

$$o_t = \sigma(W_o * x_t + U_o * out_{t-1} + b_o) \quad (4)$$

Internal state

$$state_t = a_t * i_t + f_t * state_{t-1} \quad (5)$$

Output

$$out_t = \tanh(state_t) * o_t \quad (6)$$

At each time step t, LSTM receives the previous time step output and current time step input x_t. After LSTM algorithm processing, internal $state_t$ is updated from the last $state_{t-1}$ (5), and output the processed information out_t (6). As for the state transaction, the forget gate f control the percentage of the former state in order to save (3).

After activated input data flow (1), an input gate is set to control the effects of accepting cellular memory (2), input gate layer i decided what to update and a provide the information. With updated $state_t$ (5), the output gate dominates and filter the data flow (4), output the current data (6).

III. Results & Discussion

A. Principal Component Analysis

For Recurrent Neural Network, model training is commonly limited by the fact that input features are highly correlated and the input information is redundant. PCA is necessary, in this sense, to remove the redundant features, thus help improve RNN model performance. Fig. 4 demonstrated that by applying PCA, many original input features are identified to be highly related. However, after PCA transformation only a few axes are responsible and need to be considered. For example, the first PCA feature explained over 40% variance of the data, while the 13th to 26th PCA features nearly explained nothing about the data. PCA feature 4 was an "elbow point of the curve. After PCA feature 4, no other feature explained more than 5% of the observed variance.

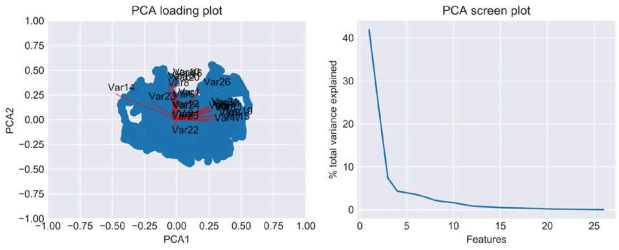

Figure 4. PCA variables illustration and all features deviation occupation (left panel), and the power of each transformed PCA axis in explaining the overall data variance (right panel).

B. Recurrent Neural Network Model

We employed LSTM implementation of RNN model, used 80% of the observed data as training data, and the rest as testing data. To perform the best model, many parameter combinations have been evaluated. The final best model consists of one layer LSTM with 15 nodes and "relu" activation function. To avoid overfitting, 20% nodes dropout was applied. Batch size for training was the length of the training data and 100 epochs were enough to achieve a relatively good model. The RNN model was applied under four conditions that with one, four, ten, and twelve PCA transformed features (Fig. 5). Considering only one PCA axis, which explained more than 40% of the observed variance (Fig. 4), RNN model showed low prediction accuracy (Fig. 5 upper left). Considering the top four transformed PCA features, RNN model performance was largely improved. For example, pearson correlation between observed and modeled energy use was as high as 0.7 (Fig. 5 upper right). Further considering two more features led to the best model that had the highest pearson correlation and the lowest loss (Fig. 5 lower left). However, after using ten input features, keep increasing input features did not benefit the RNN modeling, rather significantly degraded RNN model by introducing unnecessary redundant information (Fig. 5 lower right).

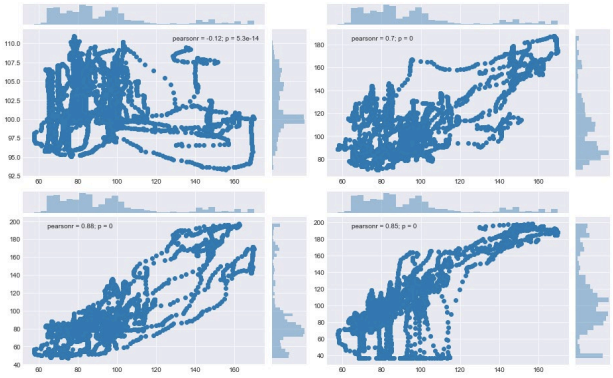

Figure 5. Scatter of test data and model predicted data with 4 conditions that considered one, four, ten, and twelve PCA features.

C. Limitation and Future Work

Although PCA removes the redundant features and generally leads to a better LSTM model, limitations also exist and need to be improved in the future experiments:

1) First, features are mostly based on one house where temperature and humidity are similar, other conditions such as seasons, building style are not considered that may influence predicting building energy use to a large extent. Collecting more features from different conditions ought to be adapt.

2) Second, the experiment only uses the data of a house as an inference, and it is difficult to map the relationship between the prediction and time series of the entire smart grid. If the prediction needs to be more accurate, multiple measurements of the buildings and the area are needed to meet the diversity, and more situations can be considered in the real prediction.

3) Third, the model is not perfect. In this model, each hidden layer uses only one layer of training, the model is relatively simple, and the activation method also uses the traditional activation equation. When adjusting the number of layers and parameters of the model, the model is more complicated when the hidden layer is set to 2, but the effect is slightly lower than that of one layer and the training time space cost is relatively high. Therefore, the optimization model can be used as a future optimization point to improve the forecast.

4) Conclusion

Implementation of smart grid in terms of efficiently deliver energy from power plant to homes and business is greatly dependent on effective prediction on building energy usage. This study aims to predict the consumption of household electricity usage by combining feature selection analysis and advanced machine learning model. Our results showed that although principal component analysis offered a great mathematical foundation to effectively transform and rank import features, the necessary number of effective feature used in machine learning model is generally unknown, thus pose a great challenge on precisely modeling and prediction future energy use. Our modeling experiment with LSTM

recurrent neural network demonstrated that model performance was significantly sensitive to the number of input features, furthermore, the model using ten input features performed best. This study provided a promising framework to improve the smart grid energy prediction implementation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Mengxuan, Jinglin and Heidan conducted the research together. Mengxuan and Jinglin analyed the data and Mengxuan did the experiments. Jinglin and Heidan raise suggestions and valuable features for the research. Mengxuan wrote the paper. Jinglin and Heidan add instructions in their field. All authors had approved the final version.

REFERENCES

[1] D. Lindley, "The energy storage problem," *Nature*, vol. 463, Jan. 2010.

[2] A. Kavousian, R. Rajagopal, and M. Fischer, "Ranking appliance energy efficiency in households: Utilizing smart meter data and energy efficiency frontiers to estimate and identify the determinants of appliance energy efficiency in residential buildings," *Energy Build*, vol. 99, pp. 220-230, 2015.

[3] K. S. Cetin, P. C. Tabares-Velasco, and A. Novoselac, "Appliance daily energy use in new residential buildings: Use profiles and variation in time-of-use," *Energy Build*, vol. 84, pp. 716-726, 2014.

[4] Z. Fan, *et al.*, "Smart grid communications: Overview of research challenges, solutions, and standardization activities," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 21-38, 2013.

[5] R. G. Pratt, C. C. Conner, B. A. Cooke, and E. E. Richman, "Metered end-use consumption and load shapes from the ELCAP residential sample of existing homes in the Pacific Northwest," *Energy Build*, vol. 19, no. 3, pp. 179-193, 1993.

[6] G. Johnson and I. Beausoleil-Morrison, "Electrical-end-use data from 23 houses sampled each minute for simulating micro-generation systems," *Appl. Therm. Eng.*, vol. 114, pp. 1449-1456, 2016.

[7] H. X. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 16, no. 6, pp. 3586-3592, 2012.

[8] N. Arghira, L. Hawarah, S. Ploix, and M. Jacomino, "Prediction of appliances energy use in smart homes," *Energy*, vol. 48, no. 1, pp. 128-134, 2012.

[9] S. H. Ling, F. H. Leung, H. Lam, and P. K. Tam, "Short-term electric load forecasting based on a neural fuzzy network," *IEEE Trans. Ind. Electron.*, vol. 50, no. 6, pp. 1305-1316, 2003.

[10] A. Veit, C. Goebel, R. Tidke, C. Doblander, and H. A. Jacobsen, "Household electricity demand forecasting: Benchmarking state-of-the-art methods," in *Proc. the 5th International Conference on Future Energy Systems*, 2014, pp. 233-234.

[11] Z. Guo, Z. J. Wang, and A. Kashani, "Home appliance load modeling from aggregated smart meter data," *IEEE Trans. Power Syst.*, vol. 30, no. 1, pp. 254-262, 2015.

[12] F. McLoughlin, A. Duffy, and M. Conlon, "Evaluation of time series techniques to characterize domestic electricity demand," *Energy*, vol. 50, pp. 120-130, 2013.

[13] L. M. Candanedo, V. Feldheim, and D. Deramaix, "Data-driven prediction models of energy use of appliances in a low-energy house," *Energy and Buildings*, vol. 140, pp. 81-97, 2017.

[14] C. L. Hor, S. J. Watson, and S. Majithia, "Analyzing the impact of weather variables, on monthly electricity demand," *IEEE Trans. Power Syst.*, vol. 20, no. 4, pp. 2078-2085, 2005.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

Mengxuan Sun was born in Shijiazhuang City, Hebei Province, China on August 17, 1995. In 2017, she graduated from the University of Essex in UK, studying master of science in big data and text analytics. She works at the artificial intelligence R & D team of China Ping An Group's life insurance headquarters as an algorithm engineer. She used to work in Chinese green grass smart technology, as a semantic algorithm programmer, to develop smart speakers. Research interests are in natural language processing, deep learning, machine learning. She won the first single model ranking of the international reading comprehension competition SQuAD2.0.

Jinglin Zhao was graduate from Hong Kong Baptist University with Bachelor of Science in Applied Psychology in 2016. She is a Course Designer worked in Xiao Hafo Kindergarten- Early Childhood Education Center in Langfang, China, which is related to designing courses based on children's different stages of development. She was interned at Data Department in Wisdom-fish Cultural Communication Co. Ltd Beijing from 2014.6 to 2015.5, which focus on Analyzing the relationship between audience's biological reaction and preference to movies by using measurement of brain activity, including Electroencephalography (EEG), Event-Related Potentials (ERP), and Galvanic Skin Responses (GSR) in Brain and Psycho-Physiology Laboratory.

Heidan Shang was born in Ankang, Shaanxi Province on 4 August 1987. He graduated from China Agricultural University at Beijing with a Bachelor's degree in Water Resources and Hydropower Engineering in 2009. In 2012, he earned his Master's degree in Water Resources and Hydropower Engineering at Northwest A&F University, located in Yangling, Shaanxi Province. In 2012, he joined in Ningxia Water Resources & Hydropower Survey Design & Research Institute Co. Ltd at Yinchuan , designing water conservancy projects. Then he worked as Product Manager in Huadong Engineering Corporation Limited.