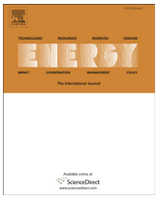
Energy 150 (2018) 49–60

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy



A novel data-driven approach for residential electricity consumption prediction based on ensemble learning

Kunlong Chen [a, b], Jiuchun Jiang [a, b, *], Fangdan Zheng [a, b], Kunjin Chen [c]

[a] *National Active Distribution Network Technology Research Center (NANTEC), Beijing Jiaotong University, Beijing, 100044, China*
[b] *Collaborative Innovation Center of Electric Vehicles in Beijing, Beijing Jiaotong University, Beijing, 100044, China*
[c] *State Key Lab of Power Systems, Department of Electrical Engineering, Tsinghua University, Beijing, 100084, China*

ARTICLE INFO

Article history:
Received 7 August 2017
Received in revised form
30 December 2017
Accepted 8 February 2018
Available online 23 February 2018

Keywords:
Household electricity consumption
Ensemble learning
Neural network
Extreme gradient boosting

ABSTRACT

With the development of smart grid as well as the electricity market, it is of increasing significance to predict the household electricity consumption. In this paper, a novel data-driven framework is proposed to predict the annual household electricity consumption using ensemble learning technique. The extreme gradient boosting forest and feedforward deep networks are served as base models. These base models are combined by ridge regression. What is more, the importances of input features are estimated. A subset of features is selected as the important features to feed into the model to increase its accuracy. A comparison of the proposed ensemble framework against classical regression models indicates that the former can reduce by 30% of the prediction error. The results of this study show that ensemble learning method can be a convenient and accurate approach to predict household electricity consumption.

1. Introduction

Studies on the residential electricity demand have been conducted since decades before for a better planning of the generation, transmission and distribution of electricity power [1]. An accurate forecasting of household electricity consumption is also of extreme significance for the contemplation of an effective energy policy to meet the current needs of the population or to anticipate its future needs. Hence, we can make more informed decisions after having a more accurate knowledge of the determinants of demand [2].

The majority of existing studies have modeled the relationship between the electricity consumption, electricity price, household income or other selected economic features by using statistical methods like linear regression. These features can be obtained relatively more easily and they are believed to be the main factors to explain the changes in the electricity consumption. For example, Alberini et al. [3] proposed a static model and a dynamic model to estimate the influence of electricity price on household electricity consumption using a generalized method of moments (GMM) estimator. Zhou et al. [4] established a linear regression model to describe the household electricity consumption by income, price and lifestyles features based on a survey data in China. Fell et al. [1] used publicly available expenditure data and utility-level consumption data from several major U.S. cities to estimate the electricity consumption. Reiss et al. [5] proposed a model concurrently addressing several inter-related difficulties posed by nonlinear pricing, heterogeneity in consumer price sensitivity, and consumption aggregation over time. The electricity consumption is modeled with nonlinear prices based on the data of California households. While these studies can provide important implications for energy policy analysis, their limitations are also obvious. Firstly, as they generally use the extended form of linear regression model to fit the data, these models sacrifice the prediction accuracy for the privilege of higher interpretability. For example, they can answer the questions like how strong the relationship between each feature and the predictive variable is or how accurately we can estimate the effect of each feature on predictive variable [6]. Nevertheless, these models are unable to capture the complex interaction of structures in the data, resulting in a lower prediction accuracy. Further, many studies are limited to geographically narrow regions, making it difficult to extrapolate their results to other areas with different climates, housing stock and electricity suppliers.

Using time series data of historical electricity consumption to predict future electricity consumption is another approach. For

[*] Corresponding author. National Active Distribution Network Technology Research Center (NANTEC), Beijing Jiaotong University, Beijing, 100044, China.
E-mail addresses: 15121396@bjtu.edu.cn (K. Chen), jcjiang@bjtu.edu.cn (J. Jiang), fdzheng@bjtu.edu.cn (F. Zheng).

https://doi.org/10.1016/j.energy.2018.02.028
0360-5442/© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

ANN	Artificial neural networks model
BM	The Blending model proposed in this article
DL	Deep Learning model
ERT	Extremely Randomized Trees model
FDN	Feedforward Deep Networks
GBDT	Gradient Boosting Decision Trees model
GMM	Generalized Method of Moments
IG	Information Gain
KNN	K-Nearest neighbors model
LR	Linear regression model
RF	Random Forests model
RMSE	Root Mean Squared Error
SVM	Support Vector Machine model
US	the United states
XGboost	Extreme gradient boosting

example, Chang et al. [7] proposed a weighted evolving fuzzy neural network approach to model the monthly residential electricity consumption. Kaytez et al. [8] compared the prediction performance of electricity consumption based on three mainstream machine learning algorithms. Dilaver et al. [9] established a structural time series model to forecast the Turkish residential electricity consumption. An et al. [10] used multi-output feedforward neural network with empirical-mode-decomposition-based signal filtering to forecast the electricity consumption. These methods can achieve high prediction accuracies. However, they suffer from a lack of interpretability. It is hard for practitioners to recognize the factors which contribute to the changes of electricity consumption. In addition, historical electricity consumption data is often gathered at the state-level or region-level. To elaborate the analysis, the household-level data is needed.

In recent years, the developments of big data techniques and machine learning algorithms lead to the wave of artificial intelligence. These advancements allow us to deal with high-dimensional data with highly complicated data structure. The aim of this paper is to introduce a state-of-the-art machine learning technique called “ensemble learning” to forecast the annual household electricity consumption. The high-dimensional raw survey data can be fed into the model without prior knowledge of feature selection. The proposed framework can predict household electricity consumption with higher accuracy than classical methods. What is more, the importance of each input feature can be calculated automatically. Hence further studies can be carried out based on the results. The main contributions of this paper are as follows.

(I) Two advanced machine learning algorithms, extreme gradient boosting and feedforward deep networks are developed as base models. A new ensemble method for the prediction of household electricity consumption is proposed to combine different base models, which is able to further improve the forecasting accuracy.
(II) Ridge regression is introduced as a combining method to form the ensemble forecast, which can decrease the chance of overfitting for the reason that base models may be co-related.
(III) Information gain is introduced as a metric to measure the importance of each feature and select a subset of informative features efficiently.

The remainder of the paper is organized as follows. In section 2, the principles of ensemble learning are introduced. The structure of the proposed ensemble learning framework is detailed as well. Section 3 presents the prediction results of the proposed method compared with those of several classical methods. In section 4, some related topics are discussed.

2. Methods

In this paper, the ensemble learning technique is introduced to combine two different models to generate a better model to predict the electricity consumption. The first part is called Extreme Gradient Boosting (XGboost) forest, which combines several different XGboost tree models [11]. The second model is called feedforward deep networks (FDN) [12]. The ensemble model takes the advantages of XGboost and FDN, both are practical techniques in dealing with complicated data, thus exhibiting a better prediction performance. An advanced regression technique called ridge regression is introduced to combine the XGboost forest and FDN. Furthermore, the mutual information is calculated to select relevant features which can fit the data better.

2.1. Extreme gradient boosting

XGBoost was recently proposed by Chen et al. [11], which combines a series of regression trees (see Appendix B) into a strong model. Although regression trees can capture complex interaction of structures in the data, they are also very noisy [13]. For example, light changes in the data can drastically change the structure of the tree [14]. Therefore in most cases, a single regression tree is inadequate for a good regression model. To get a better performance, a natural way is to combine a number of regression trees into an ensemble. Regression trees model is an excellent base learner for ensemble learning. First, separate trees can be easily added together, in the same way that individual predictors can be added together in a regression model, to generate a prediction. What is more, trees can be generated very quickly [14].

In XGboost, K additive functions are used to predict the output. That is [11]:

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathscr{F} \tag{1}$$

where $\mathscr{F} = \{f(x) = w_{q(x)}\}(q : \mathscr{R}^m \rightarrow T, w \in \mathscr{R}^T)$ is the space of regression trees as presented in Appendix A. Each f_k refers to an independent tree structure q and leaf weights w. Besides, q denotes the structure of each tree that maps an example to the corresponding leaf index and T is the number of leaves in the tree.

The final optimization target can be represented as:

$$\mathscr{L}_t = \sum_i l(\widehat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2. \tag{3}$$

l refers to a differentiable convex loss function that measures the difference between the prediction $\widehat{y}_i$ and the target y_i. In the regression task, l is normally set to squared-error function. $\Omega(f)$ is the penalty term, which avoids the problem of over-fitting in the model fitting process. As a result, a model with simple and