Applied Energy 127 (2014) 1–10

Contents lists available at ScienceDirect

Applied Energy

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

Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques

Cheng Fan, Fu Xiao *, Shengwei Wang

Department of Building Services Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong

HIGHLIGHTS

- A data mining based method is proposed to predict building energy consumption.
- The outlier detection method can identify abnormal building operating patterns.
- He recursive feature elimination technique is effective in selecting optimal inputs.
- The prediction performances of eight popular predictive algorithms are studied.
- Ensemble models built on the eight base models have the best performances.

ARTICLE INFO

Article history:
Received 28 November 2013
Received in revised form 5 March 2014
Accepted 6 April 2014
Available online 26 April 2014

Keywords:
Building energy prediction
Data mining
Feature extraction
Clustering analysis
Recursive feature elimination
Ensemble model

ABSTRACT

This paper presents a data mining (DM) based approach to developing ensemble models for predicting next-day energy consumption and peak power demand, with the aim of improving the prediction accuracy. This approach mainly consists of three steps. Firstly, outlier detection, which merges feature extraction, clustering analysis, and the generalized extreme studentized deviate (GESD), is performed to remove the abnormal daily energy consumption profiles. Secondly, the recursive feature elimination (RFE), an embedded variable selection method, is applied to select the optimal inputs to the base prediction models developed separately using eight popular predictive algorithms. The parameters of each model are then obtained through leave-group-out cross validation (LGOCV). Finally, the ensemble model is developed and the weights of the eight predictive models are optimized using genetic algorithm (GA).

The approach is adopted to analyze the large energy consumption data of the tallest building in Hong Kong. The prediction accuracies of the ensemble models measured by mean absolute percentage error (MAPE) are 2.32% and 2.85% for the next-day energy consumption and peak power demand respectively, which are evidently higher than those of individual base models. The results also show that the outlier detection method is effective in identifying the abnormal daily energy consumption profiles. The RFE process can significantly reduce the computation load while enhancing the model performance. The ensemble models are valuable for developing strategies of fault detection and diagnosis, operation optimization and interactions between buildings and smart grid.

1. Introduction

The ever-increasing energy consumption has raised worldwide concern over the issues of environmental degradation, energy security and geopolitics. According to statistics from the International Energy Agency (IEA), buildings are responsible for 32% of total final energy consumption [1]. In terms of primary energy consumption, buildings represent around 40% in most IEA countries [1]. Such figures would be much higher in less industrial-oriented districts. For example, the building sector in Hong Kong accounts for over 60% of the total final energy consumption and over 90% of electricity use [2]. Building energy efficiency is of great importance to global sustainability.

In the past two decades, researchers have devoted themselves in the improvement of building energy efficiency. Building energy consumption prediction has drawn special attention as it is often needed in developing various strategies for improving building energy performance, e.g., fault detection and diagnosis [3], demand side management for smart grid [4]. According to the time-scale of prediction, the existing research on building energy prediction can be broadly classified into three categories, i.e., short-term (i.e., up

* Corresponding author. Tel.: +852 2766 4194; fax: +852 2765 7198.
E-mail address: linda.xiao@polyu.edu.hk (F. Xiao).

http://dx.doi.org/10.1016/j.apenergy.2014.04.016
0306-2619/

to one week ahead), medium-term (i.e., from one week to one year ahead) and long-term (i.e., longer than one year ahead) predictions [5]. Currently, the short-term prediction is the main focus due to its close linkage to the day-to-day operations [6]. The short-term prediction mainly focuses on the predictions of the daily peak demand, daily energy consumption and daily load profiles [7]. Popular methods for developing prediction models include the engineering methods [8,9], statistical methods [10,11], gray-box modeling methods [12,13], machine learning and artificial intelligence methods [14–18].

Nowadays, buildings are becoming not only energy-intensive, but also information-intensive. The rich data provides convenience in modeling complex and nonlinear processes in building operations. Meanwhile, it can be tedious and time-consuming to process large-scale data. In terms of short-term building energy predictions, the existing research needs improvement in three aspects. Firstly, before the development of prediction models, efficient and effective methods should be developed to enhance the quality of massive building energy consumption data. Secondly, existing research mainly utilized domain knowledge or conventional filter methods to select input variables. Consequently, some useful knowledge may be overlooked and the resulting models may not work well under different conditions. To overcome such shortcomings, a data-driven input selection process is proposed in this study. Thirdly, individual predictive algorithms have their own pros and cons. A more advanced data mining technique, ensemble learning, can develop composite models to improve the accuracy and stability of predictions.

This paper develops an effective approach to constructing the prediction models of the next-day energy consumption and peak power demand, taking into account the three deficiencies of existing research mentioned above. Abnormal building energy consumption profiles are firstly identified and removed using feature extraction, clustering analysis, and the generalized extreme studentized deviate (GESD). Base models are then developed using eight popular predictive algorithms. A data-driven input selection algorithm, the recursive feature elimination (RFE), is applied to find inputs to the eight base models separately. The ensemble models are constructed by combining eight base models. Genetic algorithm (GA) is used to optimize the weights of eight base models in the final ensembles. The proposed approach is applied to analyze the large energy consumption data of the tallest building in Hong Kong. The performances of individual base models and the ensemble models, as well as their computation times are compared.

2. Description of data mining techniques

2.1. Research outline

Fig. 1 shows the schematic outline of the research. One year building energy consumption data, collected at 15-min intervals, are adopted for analysis. The data preparation contains three main tasks, i.e., data transformation, feature extraction, and creation of candidate input pool. The identification of abnormal building energy consumption profiles is achieved by using clustering analysis and outlier detection. The entropy-weighted k-means (EWKM) algorithm is used as the clustering algorithm and the GESD algorithm is adopted for outlier detection. The RFE is performed to select the optimal input variables for different predictive algorithms. Model parameters are optimized through leave-group-out cross validation (LGOCV). Then, GA is used to optimize the weights of eight base models in ensemble models, which output the final prediction results of the next-day daily peak power demand and daily energy consumption.

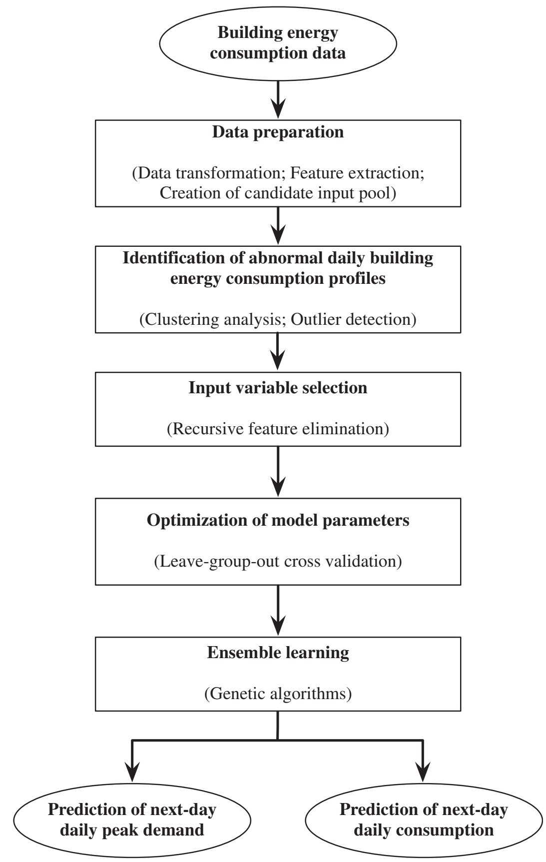

Fig. 1. Schematic outline of the research.

2.2. Clustering analysis

Clustering analysis aims to group observations with similar characteristics within the same cluster. The similarities between any pairs of observations are normally evaluated using distance-based metrics, such as the Manhattan and Euclidean metrics. The desired clustering result aims to maximize observation similarities within the same cluster and minimize the similarities between different clusters. Clustering analysis has been successfully used to preprocess large data sets, identify outliers and discover underlying patterns [20]. In this study, the entropy weighted k-means (EWKM) method is adopted to identify typical building energy consumption profiles. Three parameters should be specified to perform the EWKM algorithm, i.e., the cluster number (k), the weight distribution parameter (λ), and the convergence threshold (δ). The optimal parameter values can be determined using either internal validation methods (e.g., Davies–Bouldin index, Silhouette index and Dunn index) or external validation methods (e.g., Purity, F-measure and Normalized mutual information) [20]. The details of EWKM algorithm can be found in [19].

2.3. Generalized extreme studentized deviate (GESD)

Outliers are observations which appear to be inconsistent with the remainder of a specific data set [21]. Outliers may arise due to various reasons, such as human error, instrument error and change of system behavior. Among the existing algorithms, the GESD algorithm [22] was highly recommended because of its flexibility under various conditions [23]. It has been implemented in detecting abnormality in building energy consumption data and proved to be computationally efficient in handling large building energy data [3,24].