Renewable and Sustainable Energy Reviews 81 (2018) 1192–1205

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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

A review of data-driven building energy consumption prediction studies

Kadir Amasyali, Nora M. El-Gohary*

Dept. of Civil and Environmental Engineering, Univ. of Illinois at Urbana-Champaign, 205 N. Mathews Ave., Urbana, IL 61801, USA

ARTICLE INFO

Keywords:
Building energy
Energy consumption prediction
Data-driven methods
Machine learning

ABSTRACT

Energy is the lifeblood of modern societies. In the past decades, the world's energy consumption and associated CO_2 emissions increased rapidly due to the increases in population and comfort demands of people. Building energy consumption prediction is essential for energy planning, management, and conservation. Data-driven models provide a practical approach to energy consumption prediction. This paper offers a review of the studies that developed data-driven building energy consumption prediction models, with a particular focus on reviewing the scopes of prediction, the data properties and the data preprocessing methods used, the machine learning algorithms utilized for prediction, and the performance measures used for evaluation. Based on this review, existing research gaps are identified and future research directions in the area of data-driven building energy consumption prediction are highlighted.

1. Introduction

Buildings represent a large portion of the world's energy consumption and associated CO_2 emissions. For example, the building sector represents 39% and 40% of the energy consumption and 38% and 36% of the CO_2 emissions in the U.S. [1] and Europe [2], respectively. The use of energy that is generated from fossil fuels contributes CO_2 emissions and causes air pollution and global warming. Prediction of building energy consumption is crucial for improved decision making towards reducing energy consumption and CO_2 emissions, because it can assist in evaluating different building design alternatives and building operation strategies (in terms of their energy efficiency) and improving demand and supply management. However, building energy consumption prediction remains to be a challenging task due to the variety of factors that affect the consumption such as the physical properties of the building, the installed equipment, the outdoor weather conditions, and the energy-use behavior of the building occupants [3].

Two main approaches have been taken for building energy consumption prediction: physical modelling approach and data-driven approach. Physical models (also known as engineering methods or white-box models) rely on thermodynamic rules for detailed energy modelling and analysis. Examples of building energy simulation software that utilize physical models include EnergyPlus, eQuest, and Ecotect. These types of software calculate building energy consumption based on detailed building and environmental parameters such as building construction details; operation schedules; HVAC design information; and climate, sky, and solar/shading information [4]. However, some of such detailed data may not be available to the users at the time of simulation. Failure to provide accurate input can result in poor prediction performance.

Data-driven building energy consumption prediction modelling, on the other hand, does not perform such energy analysis or require such detailed data about the simulated building, and instead learns from historical/available data for prediction. Data-driven energy consumption prediction has gained a lot of research attention in recent years [5], despite its possible limitations (as discussed in Section 8). In response, a number of review studies on the analysis of existing data-driven approaches has been published. The reviews mostly focused on the machine learning methods/algorithms used in previous research efforts. Despite the importance of these efforts, there is still a lack of review studies that analyze existing data-driven approaches from a more multivariate perspective, including data aspects such as what data types and sizes were used and what features were selected for learning. Such a review would help reveal existing research gaps in the field of data-driven building energy consumption prediction and point towards future research directions.

To address this gap, this paper offers a review of data-driven building energy consumption prediction studies that utilized machine learning algorithms, including support vector machines (SVM), artificial neural networks (ANN), decision trees, and other statistical algorithms. The paper focuses on reviewing the types of buildings, temporal granularities, types of energy consumption predicted, types of data, types of features, and data sizes in the existing studies; and

* Corresponding author.
E-mail addresses: amasyal2@illinois.edu (K. Amasyali), gohary@illinois.edu (N.M. El-Gohary).

http://dx.doi.org/10.1016/j.rser.2017.04.095
Received 4 January 2016; Received in revised form 23 March 2017; Accepted 27 April 2017
1364-0321/

provides a discussion of the review results and future research directions. The paper is organized as follows. Section 2 provides a concise overview of existing review studies on data-driven building energy consumption prediction and identifies the gaps in this area. Section 3 gives a brief introduction on the background of data-driven approaches. Section 4 defines the methodology used in this review study. Section 5 reviews previous studies in terms of the scopes of prediction, the data properties and the data preprocessing methods used, the machine learning algorithms utilized for prediction, and the performance measures used for evaluation. Section 6 discusses the previous studies in terms of the temporal granularities of prediction, the types of buildings, and the types of energy consumption predicted. Finally, Section 7 discusses future research directions, Section 8 discusses the limitations of data-driven energy consumption prediction, and Section 9 summarizes the conclusions.

2. Existing review studies on data-driven building energy consumption prediction

Data-driven building energy consumption prediction gained a lot of attention in recent years. In response, a number of review studies has focused on the analysis of existing data-driven efforts. For example, Zhao and Magoulès [4] classified building energy consumption prediction methods as elaborate engineering methods, simplified engineering methods, statistical methods, ANN-based methods, SVM-based methods, and grey models; and conducted some comparative analysis in terms of model complexity, ease of use, running speed, inputs needed, and accuracy. Ahmad et al. [2] focused on the review of ANN-based, SVM-based, and hybrid methods and discussed the principles, advantages, and disadvantages of these methods. Fumo [6] summarized the classification of building energy consumption prediction methods proposed by various studies and placed a special emphasis on the review of model calibration and verification and weather data used for modelling. Li and Wen [7] conducted an inclusive review; they reviewed state-of-the-art studies not only on building energy modelling and prediction but also on building critical component modelling (e.g., photovoltaic power generation modelling), building energy modelling for demand response (e.g., weather condition forecasting), agent-based building energy modelling, and system identification for building energy modelling. Li et al. [8] reviewed the methods for building energy benchmarking and proposed a flowchart that intends to assist users in choosing the proper prediction method. Chalal et al. [9] focused on both building scale and urban scale energy consumption prediction and further classified and discussed the available methods within each scale. Wang and Srinivasan [10] reviewed and compared the principles, applications, advantages, and disadvantages of single AI-based methods (e.g., ANN and SVM) and ensemble methods.

The majority of these studies provided a comprehensive review on energy consumption prediction research efforts with a particular focus on the machine learning methods/algorithms used in these research studies. Despite the importance of these review efforts, there is still a lack of review studies that cover building energy consumption prediction research in terms of the scopes of prediction (e.g., heating energy consumption), the types of data used (e.g., real data, simulated data), the types of features used for prediction (e.g., outdoor weather conditions, indoor environmental conditions), the sizes of the data (e.g., duration of data collection, number of data instances), and the data preprocessing methods utilized (e.g., data reduction). Such a review is essential for identifying the research gaps and highlighting the future research directions in the field of data-driven building energy consumption prediction.

3. Background

Developing a data-driven model, typically, consists of four primary steps: data collection, data preprocessing, model training, and model testing. In the field of building energy consumption prediction, data collection involves collecting historical/available data for model training such as outdoor weather condition and electricity consumption data. Data preprocessing may include data cleaning, data integration, data transformation, and/or data reduction. Model training is the training of the model using a training dataset. Model testing aims to evaluate the model using standard evaluation measures.

SVM, ANN, decision trees, and other statistical algorithms are the most commonly-used supervised machine learning algorithms for model training. SVM is a kernel-based machine learning algorithm, which can be used for both regression and classification [11]. The algorithm is good at solving non-linear problems even with a relatively small amount of training data [4]. SVM solves a non-linear problem through transforming the non-linearity between features x_i (e.g., dry-bulb temperature and global solar radiation) and target y_i (e.g., cooling energy consumption) using linear mapping in two steps. First, it projects the non-linear problem into a high-dimensional space and determines the function f(x) that fits best in the high-dimensional space. Second, it applies a kernel function to make the complex non-linear map a linear problem. For further details on the prediction principle using SVM, the readers are referred to [9]. SVM is one of the most robust and accurate algorithms and has been listed in the top-ten most influential data mining algorithms in the research community by the IEEE International Conference on Data Mining [11]. It was found to outperform other machine learning algorithms in numerous applications. In order to increase the computational efficiency of SVM, least squares SVM (LS-SVM) (e.g., [12]) and parallel SVM (e.g., [13]) were also implemented in the field of building energy consumption prediction.

ANN is a non-linear computational model, inspired by the human brain. A typical ANN includes three sequential layers: the input layer, the hidden layer, and the output layer. Each layer has a number of interconnected neurons, and each neuron has an activation function. Three types of parameters are typically used to define ANNs: the interconnection pattern between the neurons of the different layers, the learning process of updating the weights of the interconnections, and the activation function that converts a neuron's weighted input to its output activation [14]. In ANN, each feature (e.g., dry-bulb temperature) is multiplied by its corresponding neuron weight and summed up with the bias. The activation function is then applied to determine the output (e.g., cooling energy consumption). For further details on the prediction principle using ANN, the readers are referred to [9]. ANN is one of the most popular algorithms used in building energy consumption prediction [2]. Examples of ANNs include the back propagation neural networks (BPNN), radial basis function neural networks (RBFNN), general regression neural networks (GRNN), feed forward neural network (FFNN), and adaptive network-based fuzzy inference system (ANFIS). Other methods that can be used in conjunction with ANN include the hierarchical mixture of experts (HME), fuzzy c-means (FCC), and multilayer perceptron (MLP).

Decision tree algorithms use a tree to map instances into predictions. In a decision tree model, each non-leaf node represents one feature, each branch of the tree represents a different value for a feature, and each leaf node represents a class of prediction. Decision trees is a flexible algorithm that could grow with an increased amount of training data [15]. The classification and regression trees (CART), chi-squared automatic interaction detector (CHAID), random forest (RF), and boosting trees (BT) are the most widely-used decision tree methods in the area of building energy consumption prediction.

Other statistical algorithms include multiple linear regression (MLR), general linear regression (GLR), ordinary least squares regression (OLS), autoregressive (AR), autoregressive integrated moving average (ARIMA), Bayesian regression, polynomial regression (poly), exponential regression, multivariate adaptive regression splines (MARS), case-based reasoning (CBR), and k-nearest neighbors (kNN).

Algorithms used for developing energy consumption prediction