Applied Energy 231 (2018) 331–342

Contents lists available at ScienceDirect

Applied Energy

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

Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering

Kehua Li[a], Zhenjun Ma[a,*], Duane Robinson[a], Jun Ma[b]

[a] Sustainable Buildings Research Centre, University of Wollongong, 2522, Australia
[b] SMART Infrastructure Facility, University of Wollongong, 2522, Australia

HIGHLIGHTS

- A strategy was developed to identify building typical electricity usage profiles.
- This strategy consists of intra-building clustering and inter-building clustering.
- This strategy can discover electricity usage behaviors of multiple buildings.
- This strategy outperformed two single-step clustering strategies.

ARTICLE INFO

Keywords:
Building electricity usage
Data mining
Gaussian mixture model
Hierarchical clustering
Performance comparison

ABSTRACT

This paper presents a clustering-based strategy to identify typical daily electricity usage (TDEU) profiles of multiple buildings. Different from the majority of existing clustering strategies, the proposed strategy consists of two levels of clustering, i.e. intra-building clustering and inter-building clustering. The intra-building clustering used a Gaussian mixture model-based clustering to identify the TDEU profiles of each individual building. The inter-building clustering used an agglomerative hierarchical clustering to identify the TDEU profiles of multiple buildings based on the TDEU profiles identified for each individual building through intra-building clustering. The performance of this strategy was evaluated using two-year hourly electricity consumption data collected from 40 university buildings. The results showed that this strategy can discover useful information related to building electricity usage, including typical patterns of daily electricity usage (DEU) and periodical variation of DEU. It was also shown that this proposed strategy can identify additional electricity usage patterns with a less computational cost, in comparison to two single-step clustering strategies including a Partitioning Around Medoids-based clustering strategy and a hierarchical clustering strategy. The results obtained from this study could be potentially used to assist in improving energy performance of university buildings and other types of buildings.

1. Introduction

Buildings consume about 40% of global primary energy and produce more than 30% of CO_2 emissions [1]. To improve building energy efficiency and sustainability, various technologies and solutions such as desiccant cooling [2], renewable energy integration [3], phase change materials [4], optimal control and advanced design [5] have been investigated.

Identification of building typical energy usage profiles has also been considered as a promising way to assist in understanding building energy consumption characteristics and helping the development of effective strategies to improve building energy efficiency [6]. Building energy usage profile is a time series data on energy usage of the whole building over a given period [7]. Cluster analysis, as a powerful tool which can effectively group similar objects while ensuring distinction from other grouped objects [8], has been used to identify building typical load and energy usage profiles. Miller et al. [9], for instance, proposed a method called DayFilter to detect the underlying information and identify potential areas for energy savings from building performance data and sub-system metrics. In this strategy, the building daily load profiles were first transformed into character strings using Symbolic Aggregate approXimation (SAX) and the typical daily load profiles were then identified using a *k*-means clustering method. A cluster analysis strategy to identify typical building daily load profiles

* Corresponding author.
E-mail address: zhenjun@uow.edu.au (Z. Ma).

https://doi.org/10.1016/j.apenergy.2018.09.050
Received 11 May 2018; Received in revised form 21 August 2018; Accepted 5 September 2018

Nomenclature

C	C-index
$d_{i,j}$	dissimilarity between the observations i and j
G	number of clusters
K	number of mixture components
M	distance matrix
N	number of observations in the whole dataset
N_g	number of observations in the gth cluster
N_W	number of all intra-cluster observation pairs in the whole dataset
S_{max}	sum of the N_W largest pairwise distances between observations in the whole dataset
S_{min}	sum of the N_W smallest pairwise distances between observations in the whole dataset
S_W	sum of the pairwise distances between all intra-cluster observation pairs in the whole dataset

Greek letters

μ	mean of a mixture component
ω	weight of a mixture component
$\sum$	covariance matrix of a mixture component
ϕ	probability density function of a mixture component
ψ	probability density function of a Gaussian mixture model

Subscripts

g	the gth cluster
k	the kth mixture component

using a variation-focused similarity of the load profiles was presented in Ref. [6]. The performance test of this strategy based on the hourly heating energy usage data of 19 university buildings showed that the identified typical heating energy usage profiles can provide information such as the peaks and troughs of the daily heating demand, daily high heating demand period and daily load variations, which were less likely to be revealed by the strategies only focused on the magnitude of load profiles. A shape definition language-based symbolic transformation technique was used to enhance the clustering result in terms of identifying the variations in the building energy usage [10]. Capozzoli et al. [11] proposed a general framework for identification and analysis of typical energy usage profiles, in which a hierarchical clustering with Ward's linkage method was first implemented to group the energy usage profiles. Two clustering validation indices were then used to identify the optimal number of partitions. Pan et al. [12] analysed the influence of the occupancy behaviour on the electricity load patterns of residential buildings using a k-means clustering algorithm. It was shown that the poor or elderly families tended to have a significant load shifting towards weekends while the rich and young families tended to have a fluctuating daily electricity usage curve. A clustering strategy based on a k-shape algorithm was proposed in Ref. [13] to identify building typical energy usage profiles. The hourly and weekly energy usage data from ten buildings were used to validate this strategy. It was shown that this strategy was effective in detecting building energy usage patterns and improving the forecasting accuracy of the Support Vector Regression (SVR) model. Multiple clustering algorithms such as k-means, fuzzy c-means, Self-Organizing Map (SOM), Partitioning around medoids (PAM) and hierarchical clustering have been employed to identify typical energy usage profiles of buildings. To evaluate the performance of different clustering algorithms, Panapakidis et al. [14] employed eight clustering algorithms to identify typical daily electricity usage data of 27 buildings in a university campus. A combination of SOM and k-means++ showed a better performance over the other clustering algorithms in terms of the clustering error. McLoughlin et al. [15] proposed a clustering-based strategy for creating representative electricity load profiles of residential buildings in Ireland. The k-means, k-medoids and SOM were used as the clustering techniques to characterize the diurnal, intra-daily and seasonal variations of domestic electricity demand. Tardioli et al. [16] proposed a clustering-based approach to identifying representative buildings and building groups in urban-scale datasets which contained the information about building geometries, functions and heating & domestic hot water consumption. A total of 67 clusters were identified from 13,614 buildings in a city using k-means, hierarchical clustering and PAM algorithms. It is worthwhile to note that nine validation indices were employed in this study to determine the optimal clustering result. The performance evaluation of the clustering algorithms in the aforementioned studies was mainly focused on the improvement of clustering results while the computational cost of these strategies was rarely considered.

The conventional methods become less competent for identification of typical energy usage profiles from large datasets [17,18]. Many alternative methods have been used to deal with this problem. For instance, in a number of studies, large datasets were divided into small groups based on seasons or days of the week before performing cluster analysis. In Rhodes' study [19], a k-means clustering method was used to find seasonal groups of the residential electricity usage from 103 residential buildings. The daily electricity usage (DEU) profiles of each building were first categorized into four groups according to the seasons. The means of each group were then calculated as the inputs for cluster analysis. do Carmo and Christensen [20] used a k-means clustering method to identify typical space heating profiles of single-family detached homes. One-year hourly data from 139 buildings were first categorized into groups based on weekdays and weekends as well as the intensity of building heating demand before segmentation and clustering. The results were further analyzed to investigate the correlation between household characteristics and space heating profiles using binary regression analysis. In some studies, the energy usage profiles in certain periods were aggregated before clustering. For instance, Fernandes [21] proposed a method to identify the typical natural gas consumption profiles of residential buildings. The daily gas consumption profiles from 1430 households over 243 days were clustered using a fuzzy c-means algorithm. In addition, sampling techniques such as random sampling and stratified sampling have also been considered as useful dimensionality reduction techniques to reduce the computational cost [22]. The above methods can save the computational cost of cluster analysis [8,23]. However, they also considerably reduced the resolution of the input data and some meaningful information such as the variation of daily energy usage profiles based on the days of a week, seasons or holiday timetables may also be discarded.

This paper presents a new clustering-based strategy to identify typical daily electricity usage (TDEU) profiles of multiple university buildings. Different from the majority of existing studies using a single-step clustering, the TDEU profiles in the proposed strategy were identified using a two-step clustering process (i.e. intra-building clustering and inter-building clustering) and the computational cost was a key focus during the development of this strategy. It is noted that a two-stage clustering structure was used in an early study [24] to classify electricity customers that share similar electricity usage patterns while the computational cost was not a focus. In the intra-building clustering, a Gaussian mixture model (GMM) based method was selected and used to identify the TDEU profiles of each individual building. In the inter-building clustering, an agglomerative hierarchical clustering was used to further identify the TDEU profiles of multiple buildings based on the TDEU profiles identified for each individual building. The performance of the proposed strategy was evaluated using two-year hourly building electricity usage data collected from 40 university buildings. A