Energy & Buildings 166 (2018) 284–295

Contents lists available at ScienceDirect

Energy & Buildings

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

Building energy performance assessment using volatility change based symbolic transformation and hierarchical clustering

Zhenjun Ma [a], Rui Yan [a,*], Kehua Li [a], Natasa Nord [b]

[a] Sustainable Buildings Research Centre, University of Wollongong, New South Wales, 2522, Australia
[b] Department of Energy and Process Engineering, Norwegian University of Science and Technology, Norway

ARTICLE INFO

Article history:
Received 17 August 2017
Revised 18 November 2017
Accepted 1 February 2018
Available online 14 February 2018

Keywords:
Buildings
Performance evaluation
Symbolic transformation
Visualisation
Hierarchical clustering

ABSTRACT

This paper presents the development of a symbolic transformation based strategy with interpretability and visualisation for building energy performance assessment. The strategy was developed using shape definition language based symbolic transformation and hierarchical clustering. Advanced visualisation techniques including dendrogram, heatmap and calendar view were used to assist in understanding building energy usage behaviours. A comparison of this proposed strategy with a Symbolic Aggregate approXimation (SAX) based strategy was also performed. The performance of the proposed strategy was tested and evaluated using the three-year hourly heating energy and electricity usage data of a higher education building. The result demonstrated that the proposed strategy can identify distinct building energy usage behaviours. The visualisation techniques used also assisted the information discovery process. The discovered information helped to understand building energy usage patterns. The comparison of the proposed strategy with the SAX based strategy showed that the proposed strategy outperformed the SAX based strategy for the case building tested in terms of the variations in building energy usage. This proposed strategy can also be potentially used to evaluate the operational performance of building heating, ventilation and air-conditioning (HVAC) systems.

1. Introduction

The operation of buildings and building Heating, Ventilation and Air-conditioning (HVAC) systems may suffer from various issues such as equipment malfunctions, sensor reading faults, inappropriate operating procedures, incorrectly configured control systems and equipment performance degradation [1,2]. Building energy performance assessment is therefore essential to understand building energy performance levels and timely assist in identifying the potential operational issues that may influence building energy efficiency and indoor thermal comfort.

Over the last several decades, many efforts have been made on the development of appropriate methods for effective building energy performance assessment [3]. Pang et al. [2], for instance, proposed a framework to facilitate the comparison between the building actual performance and the expected performance predicted by an EnergyPlus model. Based on a set of performance indicators, Kosai and Tan [4] developed a framework for quantitative analysis of energy performance of zero energy buildings. Yan et al. [5] developed a multi-level strategy for energy performance diagnosis of buildings with limited energy usage data available. Through a case study, Dascalaki et al. [6] concluded that building typologies can be considered as a useful tool in assessing the energy performance of residential buildings.

Data mining, as an interdisciplinary subfield of computer science, is attracting increasing attention and is now being considered as an alternative solution to address the challenges faced by conventional building energy performance assessment methods [7–12]. Gao and Malkawi [8], for instance, presented a methodology using *k*-means clustering for building energy performance benchmarking. The methodology consisted of four steps, including feature selection, cluster analysis, cluster validation and interpretation. Raatikainen et al. [9] described a method using self-organizing maps, U-matrix representation, Sammon's mapping, *k*-means clustering and Davis-Bouldin index to analyse the energy consumption of school buildings. do Carmo and Christensen [10] used *k*-means cluster analysis to identify the typical heating load profiles of Danish single-family detached homes in order to facilitate the development of cost effective demand side management solutions. The use of Partitioning Around Medoids clustering algorithm and Pearson Correlation Coefficient based dissimilarity measure to identify the typical heating load profiles of higher education buildings was presented in [11], in which the typical daily

* Corresponding author.
E-mail address: ry721@uowmail.edu.au (R. Yan).

https://doi.org/10.1016/j.enbuild.2018.02.015
0378-7788/

Table 1
Description and definition of the symbols.

Symbol	Description	Lower bound	Upper bound
Stable	Virtually no variation	−0.05	0.05
Jump	Significant increase	0.20	1.00
Up	Slight increase	0.05	0.2
Down	Slight decrease	−0.20	−0.05
Plunge	Significant decrease	−1.00	−0.20

load profiles were identified on the basis of the variation similarity. A clustering method using k-shape algorithm was used by Yang et al. [12] to identify the shape patterns of time series building energy usage data in order to improve the accuracy of forecasting models. From the above studies, it can be seen that cluster analysis is the primary data mining algorithm used in building energy performance assessment and the results showed the effectiveness of using data mining algorithms in the identification of the hidden information from the massive amount of building operational data.

In data mining strategies, data transformation is often used to transform the time series data into suitable formats to support the data mining process. Symbolic transformation is one of the common families of a time series representation approach which converts numeric time series data into symbolic forms [13]. There are two types of symbolic transformation methods, which were developed based on the means of the time segments and the volatility change, that are commonly used [13]. Symbolic Aggregate approXimation (SAX) was used by Miller et al. [14] to transform building energy usage data into alphabets to identify discords and create performance motifs. SAX was also used by Fan et al. [15] to develop a methodology for temporal knowledge discovery of big data collected from building automation systems (BASs). Based on the operational cycle of a chiller identified using a *k*-means clustering algorithm, Habib et al. [16] first transformed the operational cycles into symbols using SAX and the symbolic representation was further transformed into a bag of word representation for hierarchical clustering. The performance of an air handling unit was studied by Dedemen et al. [17], in which SAX was used to detect the frequently occurring patterns and unexpected patterns in the sensor data provided by the BAS. An extension of SAX was used by Kalluri et al. [18] to extract the features that are characteristic of individual appliance transient states in an office. From the above studies, it can be seen that SAX is the main methods used for symbolic transformation of the time series data to facilitate the data mining process.

With the wide deployment of building management systems and smart meters, a massive amount of high-resolution energy usage data from buildings can now be readily available. This provides a great opportunity to better understand building energy usage characteristics and operational performance through discovering the hidden information behind this massive amount of data. However, without advanced data analytic techniques, the valuable information underneath the massive data may not be properly extracted. This paper presents a strategy for building energy performance assessment using shape definition language based symbolic transformation and hierarchical clustering. Different from the majority of the previous studies used cluster analysis with a focus on the load magnitude for building energy performance assessment, this study used the volatility change based symbolic transformation to convert the time series data into symbolic forms and the typical building energy usage profiles were identified based on the energy usage variations. The advanced visualisation techniques including dendrogram, heatmap and calendar view were used to assist in building energy performance assessment. A comparison of the proposed strategy with a SAX based symbolic transformation strategy was also performed. The performance of this proposed strategy was tested and evaluated using the three-year hourly district heating energy and electricity usage data collected from a higher education building in Norway.

2. Development of the building energy performance assessment strategy

2.1. Outline of the proposed strategy

The outline of the proposed symbolic transformation based strategy to examine the building energy performance is presented in Fig. 1. It mainly consists of four steps, including data collection, data pre-processing, data mining, and an evaluation and interpretation of the results. The first step is the collection of building energy usage data from BASs. The collected data is then pre-processed in the second step, which consists of five main tasks including outlier removal, data segmentation, small variation segments removal, data normalisation and symbolic transformation of the time series data. In this study, the generalised Extreme Studentised Deviate (ESD) test method was used to identify and remove the outliers from the raw data as it can detect one or more outliers in a univariate data set that follows an approximately normal distribution [19]. The details of this test method can be found in [19,20]. Data segmentation is to transform the data into 24-hour segments in order to form daily load profiles. In order to identify the typical daily energy usage profiles that have distinct patterns, the segments with a small difference between the daily maximum and minimum energy usage were discarded. In this study, 5.0% of the segments with the least difference among all the daily segments were considered as the small difference and were discarded. The daily load profiles were then normalised to a range of 0–1, where 1 is the daily maximum, and 0 is the daily minimum. The last step in the data pre-processing is to transform the segments of the normalised data through the symbolic representation which will be introduced in Section 2.2.

The data mining process starts to identify the pre-defined symbols and shapes and then summarises the distribution of the symbols and shapes to provide a preliminary understanding of the building energy usage behaviour. The Dice coefficient between each pair of the daily load profiles is then calculated to determine the dissimilarity measure for clustering the daily load profiles, which will be introduced in Section 2.3. A hierarchical clustering technique is used to determine the structure and the number of the clusters with the assistance of the heatmap and dendrogram based visualisation techniques. Typical daily load profiles are then formed by calculating the mean value of all the load profiles in each cluster. The distribution of the typical daily load profiles is further plotted as a calendar view to better understand the temporal distribution of the typical daily load profiles identified.

2.2. Symbolic transformation

In this study, a volatility change based method was used to capture the variations in the building energy usage data. The normalised daily load profiles were transformed into a symbolic representation form based on the Shape Definition Language (SDL) proposed by Agrawal et al. [21]. SDL is a small language which allows a variety of queries about the shapes found in histories and has the capability for blurry matching to give the primary focus on overall shape rather than the specific details [21]. Table 1 summarises the symbols used in this study for symbolic transformation, and the corresponding description and definitions. The values used in Table 1 were determined by referring to Agrawal et al. [21]. It is worthwhile to note that these values used might not be the optimal values. In this method, the symbols were defined according to the difference between the value at the *i*th time step and the