ARTICLE IN PRESS

Renewable and Sustainable Energy Reviews xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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

A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings

Clayton Miller[a,*], Zoltán Nagy[b], Arno Schlueter[c]

[a] *Department of Building, School of Design and Environment, National University of Singapore, 117566 Singapore*
[b] *Intelligent Environments Laboratory, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, 301 E Dean Keeton St C1700, 78712 Austin, TX, USA*
[c] *ETH Zürich, Institute of Technology in Architecture (ITA), Architecture and Building Systems (A/S) Stefano-Franscini-Platz 1, Zürich, Switzerland*

ARTICLE INFO

Keywords:
Building performance analysis
Data mining
Unsupervised learning
Visual analytics
Clustering
Novelty detection
Smart meter analysis
Portfolio analysis
Review
Building controls and optimization

ABSTRACT

Measured and simulated data sources from the built environment are increasing rapidly. It is becoming normal to analyze data from hundreds, or even thousands of buildings at once. Mechanistic, manual analysis of such data sets is time-consuming and not realistic using conventional techniques. Thus, a significant body of literature has been generated using unsupervised statistical learning techniques designed to uncover structure and information quickly with fewer input parameters or metadata about the buildings collected. Further, visual analytics techniques are developed as aids in this process for a human analyst to utilize and interpret the results. This paper reviews publications that include the use of unsupervised machine learning techniques as applied to non-residential building performance control and analysis. The categories of techniques covered include clustering, novelty detection, motif and discord detection, rule extraction, and visual analytics. The publications apply these technologies in the domains of smart meters, portfolio analysis, operations and controls optimization, and anomaly detection. A discussion is included of key challenges resulting from this review, such as the need for better collaboration between several, disparate research communities and the lack of open, benchmarking data sets. Opportunities for improvement are presented including methods of reproducible research and suggestions for cross-disciplinary cooperation.

1. Introduction

The creation and consolidation of measured sensor sources from the built environment and its occupants is occurring on an unprecedented scale. The Green Button Ecosystem now enables the easy extraction of data from over 60 million buildings.[1] Advanced metering infrastructure (AMI), or smart meters, have been installed on over 58.5 million buildings in the US alone [1]. The Internet-of-Things (IoT) movement provides an array of low-cost sensors, data acquisition devices, and cloud storage. A recent study has predicted a $70–150 billion impact of IoT in offices and $200–350 billion in homes [2]. A recent press release from the White House summarizes the impact of utilities and cities in unlocking these data [3]. It announces that 18 utilities, serving more than 2.6 million customers, will provide detailed energy data by 2017. This study also suggests that such accessibility will enable improvement of energy performance in buildings by 20% by 2020. A vast majority of these raw data being generated are sub-hourly temporal data from meters and sensors.

Ruparathna et al. created a contemporary review of building performance analysis techniques for commercial and institutional buildings [4]. This review was comprehensive in capturing approaches related to technical, organizational, and behavioral changes. The majority of publications considered fall within the domains of automated fault detection and diagnostics, retrofit analysis, building benchmarking, and energy auditing. These traditional techniques focus on one building or a small, related collection of buildings, such as a campus. Many require complex characteristic data about each building, such as it geometric dimensions, building materials, the age of type of mechanical systems, and other metadata, to execute the process.

Thus, a critical issue facing the building industry is how these traditional performance analysis techniques can utilize the current explosion of detailed, temporal sources. *If one has access to raw data from thousands, or even millions, of buildings, how can analysis be scaled in a reasonable way?* Someone tasked with this type of analysis needs to extract information with less known meta-data about each building and fewer inputs into the process. In response to this question,

---

* Corresponding author.
E-mail address: clayton@nus.edu.sg (C. Miller).
[1] According to: http://www.greenbuttondata.org/

http://dx.doi.org/10.1016/j.rser.2017.05.124
Received 25 May 2016; Received in revised form 19 December 2016; Accepted 19 May 2017

Please cite this article as: Miller, C., Renewable and Sustainable Energy Reviews (2017), http://dx.doi.org/10.1016/j.rser.2017.05.124

researchers from several domains have developed methods of extracting insight from raw, unlabeled data from the built environment. Often these methods fall into the category of unsupervised statistical learning. Methods from this sub-domain of machine learning are advantageous due to their ability to characterize measured or simulated performance data quickly with less analyst intervention, metadata, and ground truth labeled data. They are often used in conjunction with visual analytics approaches as the first pass in the analysis of a large dataset.

A major challenge in utilizing this body of knowledge is that unsupervised machine learning techniques have been proposed, tested, and applied concurrently in several research domains and subdomains. This situation has resulted in parallel, and mostly non-overlapping, collections of work with similar objectives, yet different publication venues and terminology. This review seeks to quantify this situation through an overview of the use of unsupervised learning techniques developed for the building industry. This task is done more holistically than previous reviews by combining key publications from a larger set of category applications and research domains. This overview is the first to specifically target unsupervised learning and visual analytics techniques for this context. The research sectors of building energy analysis, building simulation, computer science and electrical engineering are analyzed. This paper compiles a set of 100 publications released since 2005 that focus on the use of unsupervised statistical learning techniques on temporal or spatial data from non-residential buildings.

The paper is structured as follows, Section 2 gives background on the unsupervised machine learning used in the publications. Section 3 provides an overview of various trends and statistics about the publications, including specific methods, the yearly timeline of release, and which research domains and specific journal and conference publications contain them. Sections 4- Sections 7 evaluate each of the publication categories in-depth. Finally, Section 8 provides a discussion of the challenges observed and potential opportunities discovered through the review process.

1.1. *Previous reviews*

Various reviews have been completed that overlap with a part of the scope of this paper. Most of them are designed to focus on a single core domain of research; the main two areas are building operations analysis and smart grid optimization. One of the earliest reviews of artificial intelligence techniques for buildings was completed in 2003 by Krarti and covered both supervised and unsupervised methods [5]. Dounis updated this work and focused on outlining specific techniques in detail [6]. Reddy's seminal book about a large variety of analysis techniques for energy engineers includes chapters on clustering and unsupervised methods specifically [7]. Lee et al. describe a variety of retrofit analysis toolkits which incorporate unsupervised and visual analytics approaches in a practical sense [8]. Ioannidis et al. created a large ontology of data mining and visual analytics for building performance analysis, however with a strong focus on the techniques and not examples of works using them [9]. From the utility and power grid side, Morais et al. created a general overview of various data mining techniques as focused on power distribution systems [10]. Chicco covered clustering methods specifically focused on load profiling tasks [11]. Zhou et al. included the concept of customer load classification [12]. Finally, in the most recent works, Yu et al. discuss the in-depth role of data mining in the creation of energy-efficient cities and Zhao et al. cover data mining from the grid to building scale from a high-level [13,14].

2. Background of unsupervised learning and visual analytics techniques

This section gives an overview of the five categories of unsupervised learning and visual analytics techniques discussed in this paper: clustering, novelty detection, motif and discord detection, rule extraction, and visual analytics. The general nature of the techniques is discussed, and the specific methods are briefly listed along with the abbreviations that are used in Tables 1–4.

Statistical learning can be divided into two major categories: supervised learning and unsupervised learning. According to the formative text on the subject by James et al. and Hastie et al., supervised statistical learning techniques involve a set of observations x_i, $i = 1, \ldots, n$ and an associated response variable y_i [15,16]. The goal of a supervised process is to predict y_i using the features generated from x_i; this objective is straightforward to verify according to the accuracy of the algorithm in predicting that response. Unsupervised learning, in contrast, lacks the associated response y_i because it seeks merely to understand the relationship between the observations in x_i, generally in an exploratory fashion. These references go on to explain that unsupervised techniques are, thus, often more subjective in their application and are considered more challenging in their utilization in practical applications [17,15,16]. Despite the challenge, various instances of their implementation are found in the literature with relation to building performance data as outlined in the subsequent subsections.

2.1. *Clustering*

Clustering is the most common general unsupervised approach applied to building performance data. It is used automatically to generate *subgroups* of similar types of observations. James et al. describe this process as the grouping of n observations into K groups, or clusters, according to a set of generated p features [15]. The two most common types of clustering are K-means and Hierarchical clustering. A wider array of techniques has been developed to optimize the objectives of separating subgroups in more effective and efficient ways. An example of the clustering of daily energy consumption profiles can be seen in Fig. 1. This example shows three and half years of hourly energy data for a building in which the diurnal patterns are grouped

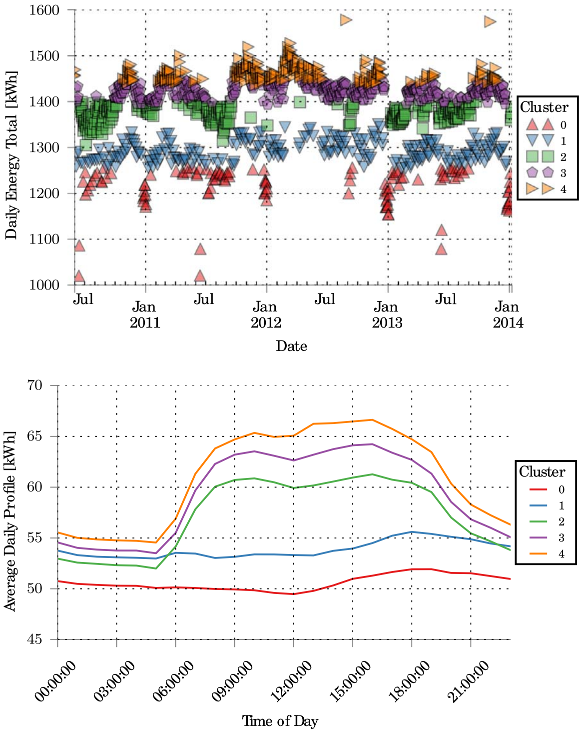

Fig. 1. Load profiles clusters from a building electricity consumption data set (adapted from [18]).