G Model
SCS-348; No. of Pages 6
ARTICLE IN PRESS
Sustainable Cities and Society xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Sustainable Cities and Society

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

Engineering advance

# Advances and challenges in building engineering and data mining applications for energy-efficient communities☆

Zhun (Jerry) Yu[a], Fariborz Haghighat[b,*], Benjamin C.M. Fung[c]

[a] College of Civil Engineering, Hunan University, Changsha 410082, Hunan, PR China
[b] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, QC, Canada H3G 1M8
[c] School of Information Studies, McGill University, Montreal, QC, Canada H3A 1X1

ARTICLE INFO

*Article history:*
Available online xxx

*Keywords:*
Data Mining
Building Energy Use
Occupant Behavior
Big Data
Review

ABSTRACT

The rapidly growing and gigantic body of stored data in the building field, coupled with the need for data analysis, has generated an urgent need for powerful tools that can extract hidden but useful knowledge of building performance improvement from large data sets. As an emerging subfield of computer science, data mining technologies suit this need well and have been proposed for relevant knowledge discovery in the past several years. Aimed to highlight recent advances, this paper provides an overview of the studies undertaking the two main data mining tasks (i.e. predictive tasks and descriptive tasks) in the building field. Based on the overview, major challenges and future research trends are also discussed.


**Contents**



## 1. Introduction

Currently the importance of improving building energy performance for saving energy and enhancing building sustainability has been widely recognized. One effective way of achieving this objective is to uncover and extract useful knowledge from building operational data (e.g. temperature, flow rate, power and equipment states) that contains abundant valuable information on actual building performance. The widespread use of *building automation systems* (BASs) enables a tremendous amount of building operational data to be stored in building databases that also continue to expand. This rapidly growing and gigantic body of stored data, coupled with the need for data analysis, has generated an urgent need for powerful tools that can extract hidden but useful knowledge from large building databases.

As an emerging and promising technology, data mining (DM) is a powerful and versatile tool to automatically extract the valuable knowledge embedded in huge amounts of data. It can be defined in many different ways. As defined by Cabena, Hadjinian,

☆ This is an Engineering Advance paper.
* Corresponding author.
*E-mail address:* haghi@bcee.concordia.ca (F. Haghighat).

http://dx.doi.org/10.1016/j.scs.2015.12.001

Stadler, Verhees, and Zanasi (1998), DM is "an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases." In the past several decades, researchers have been vigorously and successfully applying DM in many scientific, medical, and application domains such as banking, bioinformatics and new materials identification. Recently it has also been introduced into the building field that is a well-fit application area for DM, since it generates and collects vast amounts of data on system operation, occupant behavior, power consumption, climatic conditions and etc.

In general, DM includes six categories of widely accepted and implemented techniques (Usama, Gregory, & Smyth, 1996):

- Data classification (e.g. the decision tree method, support vector machine (SVM) and artificial neural network (ANN)),
- Clustering analysis,
- Association Rule Mining (ARM),
- Regression,
- Summarization, and
- Anomaly Detection.

Readers may refer to (Jia, Kamber, & Pei, 2012) for detailed information of these techniques. These techniques can be broadly categorized into predictive tasks and descriptive tasks (Jia et al., 2012). This paper reports an overview of the recent studies undertaking the two tasks in the building field.

## 2. DM applications in the building field

### 2.1. Predictive tasks

#### 2.1.1. Building energy demand prediction

The prediction of building energy demand plays an important role in improving building performance. An accurate prediction needs to take various significant influencing factors of building energy demand into consideration, such as weather conditions, HVAC equipment, building envelopes and occupant behavior. The complexity and uncertainty of these factors further adds difficulties to improve prediction accuracy. DM techniques may make a breakthrough in dealing with this complexity and uncertainty.

Zhao and Magoulès (2012) indicated that DM techniques are very applicable to building energy demand prediction since they can deal with non-linear problems. ANN and SVM are the two most widely used DM techniques for this application (Ahmad et al., 2014). Kumar, Aggarwal, and Sharma (2013) applied various ANN methods, including back propagation, recurrent ANN, auto associative ANN and general regression ANN. The adopted ANN architecture significantly influences the coefficient of variation that ranges from 2% to 40%. They concluded that ANN is more suitable for the prediction of a large set of parameters than any statistical techniques. Li, Ren, and Meng (2010) reported that in many cases SVM shows higher prediction accuracy than ANN. However, training SVM can be a very slow process due to large volume of training data. The usage of parallel SVM (Zhao & Magoulès, 2011) might be a feasible alternative.

Both ANN and SVM models operate like a "black box", meaning that the model can provide a prediction but cannot provide a justification for supporting the prediction. In order to overcome this limitation, Yu, Haghighat, Fung, and Yoshino (2010) developed a building energy demand predictive model based on the decision tree method. Its competitive advantage lies in the ability to generate accurate predictive models (92% accuracy in their study) with interpretable flowchart-like tree structures that enable users to quickly extract useful information. However, the decision tree method is basically developed for predicting categorical variables other than for predicting numerical variables.

Considering that different DM techniques have their own specialties, strengths and weaknesses, it is not easy to find the best candidate for the building energy demand prediction. The recent research trend is towards the integration of different DM techniques for more accurate prediction. For example, Chou and Bui (2014) developed an ensemble model by combining ANN and SVM to predict cooling and heating demand. In order to predict the next-day energy consumption and peak power demand, Fan, Xiao, and Wang (2014) developed ensemble models by combining eight base DM models and assigning each DM modal an optimized weight based on genetic algorithm (GA). The results show that the ensemble models achieve higher prediction accuracy than those of individual base models.

#### 2.1.2. Building occupancy and occupant behavior

Building occupancy and occupant behavior are recognized as crucial factors influencing the discrepancy between practical and simulated building energy consumption. However, it is difficult to investigate them analytically and then to develop reliable prediction models due to their complicated characteristics and stochastic nature (Yu, Haghighat, Fung, Morofsky, & Yoshino, 2011a). To meet this challenge, different stochastic models (e.g. using probabilistic methods (Sun, Yan, Hong, & Guo, 2014; Stoppel & Leite, 2014) and the Markov Chain method (Muratori, Roberts, Sioshansi, Marano, & Rizzoni, 2013)) have been proposed. However, the proposed models are severely restrained by the limitation: The modeling process tends to be complex and advanced mathematical knowledge is required.

In order to remove the above limitation, researchers have attempted to establish DM-based models. For example, Basu, Hawarah, Arghira, Joumaa, & Ploix, 2013) developed a decision-tree based model for predicting occupant behavior at the appliance level in residential buildings. D'Oca and Hong (2015) proposed a DM methodology to model office occupancy patterns and working user profiles based on big data streams. They found that the decision tree method is suitable for predicting the occupancy presence, supported by Zhao, Lasternas, Lam, Yun, and Loftness (2014), who built the occupant behavior prediction models based on appliance power consumption data in a medium-size office building. The results of both studies indicate that the modeling accuracy is very satisfactory. However, the developed decision-tree based models are static models that are hard to simulate the dynamic nature of building occupancy and occupant behavior.

The main focus of future research should be placed on testing and comparing other dynamic DM-based models and integrating them into building energy modeling programs like TRNSYS and Energy Plus. In addition, more research needs to be conducted so that architects and designers will benefit from bridging the gap between actual and predicted building energy performance.

#### 2.1.3. Fault detection diagnostics (FDD) for building systems

Automating the process of detecting equipment and system malfunctions and making a proper diagnosis can help to ensure stable or optimal building operation. In terms of the approach to formulating the diagnostics, FDD methods can be categorized as model-based methods, which are based on prior knowledge of underlying system physics, and data-driven methods, which are based on historical data (Katipamula & Brambley, 2005).

The requirement of prior knowledge, together with complex modeling processes and heavy computational burden, imposes severe constraints on the application of the model-based methods. Comparatively, the data driven methods are much easier to use since the models are normally automatically generated. Various DM techniques have been employed as data driven methods for FDD.