BUILD SIMUL
https://doi.org/10.1007/s12273-019-0599-0

Data mining algorithm and framework for identifying HVAC control strategies in large commercial buildings

Zhe Chen[1], Peng Xu[1] (✉), Fan Feng[2], Yifan Qiao[1], Wei Luo[1]

1. Department of Mechanical and Energy Engineering, Tongji University, Shanghai, 201804, China
2. Department of Mechanical Engineering, The University of Alabama, Tuscaloosa, 35487, AL, USA

Abstract

Heating, Ventilation, and Air-Conditioning (HVAC) control strategies are set arbitrarily in many commercial buildings by operators, who sometimes lack relevant skills and professional training. It is acknowledged that improving the control strategy of HVAC is feasible and valid, which as a consequence can improve the overall HVAC performance of existing buildings. However, it is quite difficult for an outsiders or a commissioning agent to tell what the HVAC control strategies are and whether they are implemented appropriately in existing buildings. This paper is intended to carry out analysis on the data about Building Automation System (BAS), as well as the data about building energy, for the purpose of identifying the control strategies of HVAC in a given building by using data mining algorithm. Then the results can be adopted by us to determine whether the building is under faulty operation or is running under suboptimal conditions. In this paper, what are proposed are algorithms of data mining identification for some specific HVAC control strategies, including DR on/off strategy, DR reset strategy and temperature reset strategy of chilled water. On the basis of data mining algorithms, a framework is then developed so as to identify these strategies, and the main scenario of this identification framework is known as analyzing many commercial buildings on an energy monitoring platform of a public building. This framework takes the sensor data obtained from HVAC, including temperature, flowrate, and electricity usage, as input, which is followed by the application of Image Segmentation and PCA algorithm for preprocessing. Then, based on these input variables, XGBoost algorithm is employed to determine whether these strategies have been implemented in buildings or not. In order to get the data for training and testing the framework, EnergyPlus Runtime Language is adopted for the application of different strategies. t is finally shown by the result that the identification algorithm can achieve the accuracy rate of 92.5% in the case studies by using one-day operation data, and the identification algorithm can arrive at the accuracy rate of 100% by using three-day operation data.

Keywords

HVAC,
control strategy,
identification,
data mining

Article History

Received: 16 August 2019
Revised: 09 November 2019
Accepted: 18 November 2019

1 Introduction

Generally speaking, commercial buildings fall within the categories of office buildings, shopping malls, hotels, cultural facilities, medical facilities, and sports facilities etc. In 2018, the total area of commercial buildings in America was 91.7 billion square feet and the Energy Use Intensity (EUI) was 199.1 thousand Btu per square foot. Apart from that, the energy consumption of American commercial buildings in 2018 has expanded to as large as 18606.9 trillion Btu, which accounted for about 18% of the total energy use (EIA 2018). With regard to this tendency, it is reported by EIA that energy consumption will grow by an average of 2.1% per year from 2012 to 2020 in the entire world in commercial sector (EIA 2018).

HVAC systems are regarded as one of the most significant parts of energy consumers in buildings. Apart from its design, the operation and control are also of great significance in terms of the final energy efficiency. It is roughly estimated that 25%–50% of the energy use in commercial buildings in

E-mail: xupeng@tongji.edu.cn

the UK has been wasted because of the improper operating conditions for the reason that both the operation and control are highly depended on the skills and motivation of its onsite operator (CIBSE 2000). It has been shown by relevant studies that poorly maintained, degraded and improperly controlled equipment is estimated to waste about a 15%–30% energy use in commercial buildings (Katipamula and Brambley 2005).

The control of HVAC system can be divided into two control patterns: one is local control at the bottom level and the other is global control, or supervisory control at the top level. As far as local control is concerned, it takes control of a setting point through an actuator (ASHRAE 2011). For instance, the temperature of cooled air after going through cooling coil is controlled by valves in supply water loop. However, setting points of the system and other related operation patterns are designated in global control, which sometimes is also named as supervisory control. Global control can be further decomposed into the control based on setting points, such as chilled water temperature control, on the basis of operation patterns, such as chiller sequence control. Based on its control logic, global control can be either established on the basis of timetable or real-time feedback (according to weather condition, building condition, electricity price, etc.).

Actually, the efficiency of HVAC system can also be improved through two-level control. It is known that the regulation over local control not only can improve comfort and reduce oscillations, but also can prolong lifetime of the equipment. In global control, the energy efficiency can be boosted by the optimization conducted for the setting points and operation mode. In this research, unless specifically noted, the control strategies concerned in the following part refers to global control strategies.

Nowadays, more and more commercial buildings are equipped with monitoring systems for energy consumption, and therefore real-time data could be acquired through the municipal platform. However, it is usually difficult for the platform to obtain the control strategies of the buildings and it is also not easy to keep track of the control strategies adopted by a variety of buildings with different control platforms. Moreover, it is shown by our site survey that maintainers are able to record only little information about the operation status of the buildings in many buildings, which accordingly provides us the precious opportunity to identify operation strategies adopted by HVAC systems in commercial buildings. For instance, in a commercial building surveyed by us, a total of 192 data collecting ports are available and the data is collected and saved every 15 minutes. In this way, the number of 35,040 pieces of time sequence data is collected in a year and the data can be employed to figure out what kinds of control strategies are truly carried out in this building. This paper attaches great importance to the development of algorithm for the purpose of extracting control strategy on the basis of these data by the application of data-mining algorithm.

Data-mining (DM) is deemed as an integral branch of computer science that can be employed to a big dataset for the aim of extracting meaningful knowledge automatically or semi-automatically. So far, this technology has been adopted in many domains from research to business (Witten and Frank 2005). For instance, Mirzaei and Reza (2012) conducted relevant analysis on the shopping record in a supermarket by applying the algorithm of association rule learning, and finally found that a certain goods were often bought together. In 2009, Google successfully predicted the influenza epidemics in winter through the application of data-mining algorithm (Ginsberg et al. 2009). In terms of studies related to buildings, the implementation of DM has become more and more prevalent, especially in the tasks of forecasting and control strategy recognition. In recent decades, the technology of DM has been widely employed to forecast energy consumption of buildings. There are actually many sorts of forecasting models, such as regression model (Wang et al. 2015; Bauer and Scartezzini 1998), SVM (support vector machine) model (Hong 2009; Niu et al. 2010; Li et al. 2009), ANN (artificial neural network) model (Kusiak et al. 2010; Kalogirou et al. 1997; Yu et al. 2010), and decision tree (Yu et al. 2010). Up till now, these forecasting models have been found in practical application in the fields of demand response, and forecasting control, etc.

Different from traditional work relevant to forecasting, identification conducted for operation strategies refers to the extraction of information about operation behaviors instead of just attaching importance to energy consumption obtained from the operation data. For instance, D'Oca and Hong (2014) established a model for the pattern of preference for window opening and closing in office buildings, and it is shown by the results obtained that those occupants who have the preference to open the window for a short period of time tend to have a small opening angle. However, the research on control identification is still in the infant stage. Yu et al. (2012) carried out related research on the statistic pattern of the operation data for the aim of understanding the control logic of the building better. More than that, many "if-then" rules were acquired by ARM (Association Rule Mining) so as to identify the bad control strategies. Fan et al. (2015) conducted similar studies on the basis of QARM (Quantitative Association Rule Mining), and this study focuses on the association rules among the time series. With the purpose of handling a huge amount of data, SAX (Symbolic Aggregate Approximation) was applied to preprocess raw data. Yu et al. (2012) also employed the same idea of ARM in their studies, but it was based on the

top level rather than concentrating on a specific subsystem or control logic. However, the disadvantage lies in the fact that the depth of data mining is not deep enough for the reason that the DM process is so general that it is quite easy to become aimless. Moreover, there are some scholars who have carried out relevant researches on the application of DM in a specific system. Motta Cabrera and Zareipour (2013) dedicated to the identification of the control strategies for lighting system by ARM in educational institutes and it is shown by the results obtained from simulation that the electricity waste in lighting is estimated to be as much as 70%. Li et al. (2017) integrated ARM with cluster algorithm for the purpose of obtaining the energy consumption pattern of HVAC system and finally came to the conclusion that high energy consumption of HVAC system would lead to the high frequency of compressors. D'Oca and Hong (2015) carried out related studies on a set of occupancy data, and four typical working user profile schedules and their distributions were obtained by employing decision tree model and cluster analysis. Actually, the DM framework proposed can be applied to a variety of different data sets, but the limitation lies in that the user profiles and patterns of occupancy are circumstantial to the given data set and the occupancy patterns are not multiple enough. Moreover, the accuracy about the implementation of the framework on other data sets hasn't been proved yet.

Apart from aiming at extracting the information related to operation behaviors, there are also some researches into the task about obtaining a better DM algorithm at the local level. For instance, Li et al. (2014) carried out analysis on the relationship between the cyclic features for the operation of furnace and the parameters of weather. Feng et al. (2017) and Qiu et al. (2019) also developed the algorithms for identifying the operation strategies such as sequence control and coordinated control with DM algorithms in a local control.

The researches mentioned above are almost all that we could find concerning the identification of control strategies. However, it is acknowledged that most of them are not established on the basis of analysis at the local level. For the purpose of better utilizing the operation data at a system level, we are required to develop more applicable framework and algorithm for identification. Different from the research mentioned previously, the method proposed by us can extract specific control strategies (chiller reset for instance). Considering that an increasing number of commercial buildings is taking the initiative to participate in set points setback and chillers oscillation under electrical Demand Response (DR) controls, an algorithm is also proposed by us to identify whether the building gets involved in the demand response project or not.

2 Method

This paper is actually aimed at developing a framework which can be employed for the identification of specific control strategies as shown in the left column of Fig. 1. The identification method is firstly developed in this section. In Section 3 a typical building is introduced to train the framework and test the trained framework of a practical office building, which is shown in the middle and right columns of Fig. 1. With regard to Section 3, it mainly attaches importance to the development of the model so as to identify specific control strategies introduced, and then how the model can be applied to a practical case is presented as well. Finally, the discussion over the results obtained from model training and case are specified in Section 4.

2.1 Specific control strategies

A total of 3 different global control strategies are chosen by us, and they are namely DR on/off strategy, DR reset

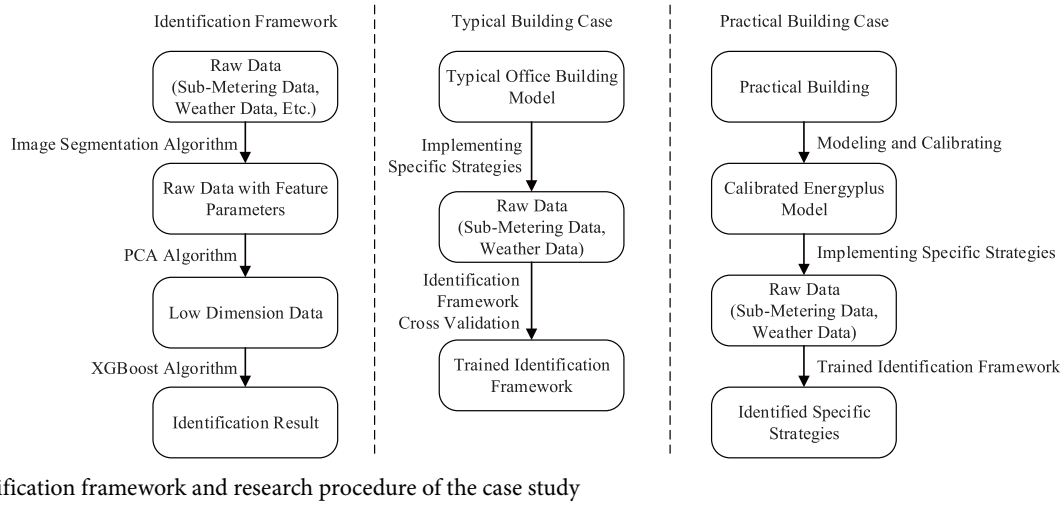

Fig. 1 Identification framework and research procedure of the case study

strategy, and reset strategy of chilled water. The first two strategies are actually tailored for the use of DR which is adopted when the demand peak occurs through employing some regulation methods, and therefore the burden of the gird can be relieved to certain extent (Li et al. 2016). Besides, as a specific control strategy, the reset method of ASHRAE 90.1 (ASHRAE 2013) was also included in the research.

2.1.1 *DR on/off strategy*

After a building's receiving a DR signal, the on/off control can be applied to chillers so as to decrease the energy consumption. The specific strategy employed in our research is that all chillers are shut down with only the chilled water loop running. Therefore, it is known that the cooling load is partly undertaken by the chilled water in the chilled water loop, as well as the thermal mass (building envelop, furniture, etc.) in the building. For the purpose of not impairing the indoor comfort, the highest acceptable temperature is set. When the indoor temperature is higher than the point set previously, chillers are activated accordingly, as is depicted in detail in Fig. 2.

2.1.2 *DR reset strategy*

Chilled water reset is able to adjust the temperature of chilled water, which as a consequence can decrease the electricity consumption and then improve the COP of the chiller plant as well. For instance, as shown in Fig. 3, the original temperature of the supply is 6.7 °C and the setpoint increases to 10 °C, when the system receives the DR signal. At the end of DR event, the setpoint temperature is reset back to 6.7 °C again.

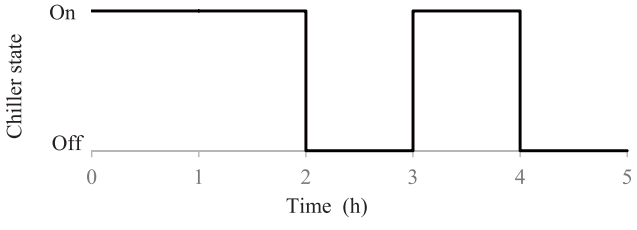

Fig. 2 Chiller state during DR period applying DR on/off strategy

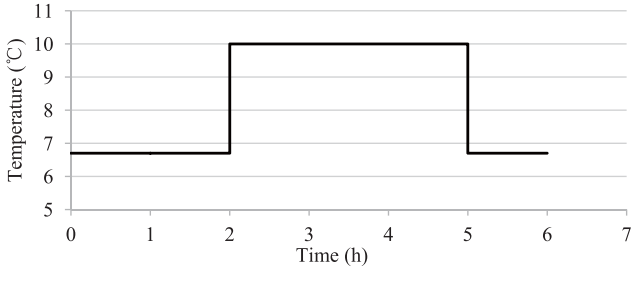

Fig. 3 Temperature reset of chilled water supply during DR period

2.1.3 *Strategy of chilled water reset*

Resetting the temperature of chilled water in accordance with the outdoor dry-bulb temperature is deemed as an effective energy-efficient approach. The specific strategy can be expressed by Eq. (1) and it is depicted in Fig. 4. It is easy for us to note from Eq. (1) that the setpoint of chilled water has a linear relationship with outdoor dry-bulb temperature and, as a matter of fact, the energy consumption of chiller plant could be approximately expressed by a quadratic equation about the temperature of chilled water supply with a factor of outdoor dry-bulb temperature (Braun 1989).

$$T_{\mathrm{chw,sup}} = \begin{cases} 7\,^{\circ}\mathrm{C} & \text{if } T_{\mathrm{out,air,db}} > 27\,^{\circ}\mathrm{C} \\ 12 - \frac{5}{11}\left(T_{\mathrm{out,air,db}} - 16\right)^{\circ}\mathrm{C} & \text{if } 16\,^{\circ}\mathrm{C} < T_{\mathrm{out,air,db}} \le 27\,^{\circ}\mathrm{C} \\ 12\,^{\circ}\mathrm{C} & \text{if } T_{\mathrm{out,air,db}} < 16\,^{\circ}\mathrm{C} \end{cases} \tag{1}$$

where $T_{\mathrm{chw,sup}}$ refers to the temperature of supply chilled water and $T_{\mathrm{out,air.db}}$ denotes the dry-bulb temperature of air outdoors.

2.2 Data description

The algorithm is mainly employed for the aim of identifying the control strategy that is applied to the building on the basis of the data collected. The variables employed in this algorithm are listed in detail in Table 1. It is easy to collect

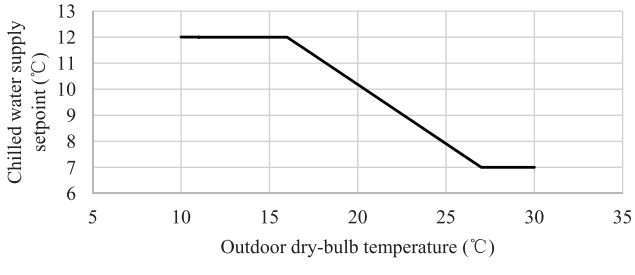

Fig. 4 Schedule of chilled water reset from ASHRAE 90.1

Table 1 Variables for identification

Parameter	Symbol
Chilled water supply temperature (°C)	$T_{\mathrm{chw,sup},i}$
Chilled water return temperature (°C)	$T_{\mathrm{chw,ret},i}$
Outdoor dry-bulb temperature (°C)	$T_{\mathrm{out,air,db},i}$
Power of the chillers (kW)	P_i
Indoor dry-bulb temperature (°C)	$T_{\mathrm{in,air,db},i}$
Chilled water primary side flow (m^3/h)	$T_{\mathrm{chw,pri},i}$
Chilled water secondary side flow (m^3/h)	$T_{\mathrm{chw,sec},i}$
Chilled water bypass flow (m^3/h)	$T_{\mathrm{chw,bypass},i}$

these variables from the automation system of building. In this research, input data is collected in a day (24 h) with an interval of 10 min.

Therefore, the final input X of the algorithm is specified as follows:

$$\begin{aligned} X = \{&T_{\text{chw,sup},1}, T_{\text{chw,sup},2}, \ldots, T_{\text{chw,sup},n}, \\ &T_{\text{chw,ret},1}, T_{\text{chw,ret},2}, \ldots, T_{\text{chw,ret},n}, \\ &T_{\text{out,air,db},1}, .T_{\text{out,air,db},2} \ldots T_{\text{out,air,db},n}, \\ &P_1, P_2 \ldots, P_n, \\ &T_{\text{in,air,db},1}, T_{\text{in,air,db},2}, \ldots, T_{\text{in,air,db},n}, \\ &F_{\text{chw,pri},1}, F_{\text{chw,pri},2}, \ldots, F_{\text{chw,pri},n}, \\ &F_{\text{chw,sec},1}, F_{\text{chw,sec},2}, \ldots, F_{\text{chw,sec},n}, \\ &F_{\text{chw,bypass},1}, F_{\text{chw,bypass},2}, \ldots, F_{\text{chw,bypass},n}\} \end{aligned} \tag{2}$$

The output is expressed as y, and the possible value is [1,2,3,0], in which 1 refers to the fact that on/off strategy is applied during DR period, 2 denotes reset strategy during DR period, 3 indicates ASHRAE reset strategy and 0 represents the situation that no one of the above-mentioned strategies is applied.

On the basis of the description mentioned above, it is easy for us to obtain some features of this problem:

1) This is a high-dimension problem. A total of 8 parameters are available and each parameter can produce 144 measurement points in 24 h and thus we are able to obtain an 1152-dimension input. However, we can only obtain about 90 sets of observation data in a cooling season (the original model without any specific control strategies is included) as far as a building model is concerned, and we can obtain about 360 sets of data if three specific control strategies are applied to the model.
2) This is a problem related to multi-classification. The output is a discrete value that can be addressed through Logistic Regression or Classification Decision Tree.
3) Every variable in X is a time series that can be employed with the purpose of improving the accuracy of identification.

2.3 Extract feature parameter

It is observed that the data obtained from the specific control strategies shows a pattern that is different from the normal pattern. Therefore, the image segmentation algorithm is adopted by us so as to separate every time-series data (Rafael et al. 2009). In the field of computer vision, a digital image is usually separated by image segmentation into several segments (also named pixel group or superpixel) according to its lines and curves. The process of image segmentation is typically region-based or edge-based (Morar et al. 2012; Pal and Pal 1993). In terms of the edge-based process that is adopted in this paper, it is mainly responsible for detecting points or lines that change rapidly in comparison with that of the neighborhood and then separating the image in accordance with these geometrical elements. The work of detection is realized through derivative operators such as Roberts operator (Dony and Wesolkowski 1999), Prewitt operator (Wang and Zhou 2008) and Sobel operator (Kanopoulos et al. 1988).

Two feature parameters, N_{zones}, and $\text{Std}_{\text{zones}}$, are defined to assess the result of segmentation. N_{zones} refers to the number of the segments and $\text{Std}_{\text{zones}}$ denotes the mean of the standard deviation of these segments, as shown in the equation as follows:

$$\text{Std}_{\text{zones}} = \frac{1}{N_{\text{zones}}} \sum_i \text{Std}_i \tag{3}$$

where Std_i denotes the standard deviation of the i^{th} segment.

For instance, Fig. 5 shows a sequence of data and N_{zones} = 7 and $\text{Std}_{\text{zones}}$ = 0.0428 are obtained after the segmentation process. Then, it is also obtained that there are 16 additional input and the new input has 1168 dimensions.

2.4 Dimension reduction

PCA (Principal Component Analysis) was employed with the aim of reducing the dimension of the data. PCA is able to transform a set of n-observations of p-dimension variables (X_1, X_2, ..., X_p) into orthogonal principal components (Z_1, Z_2, ..., Z_p) which can satisfy the requirements of Eqs. (4)–(6) (i.e., these orthogonal components are deemed as a new linear combination of the original one, with the former components having larger variance while the latter components having smaller variance). The original data has a dimension of p and the principal component has a new dimension of p' ($\leq p$) by using the first i principal components.

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p \tag{4}$$

$$\text{Var}(Z_1) > \text{Var}(Z_2) > \ldots > \text{Var}(Z_p) \tag{5}$$

$$\text{Cov}(Z_i, Z_k) = 0 \quad \text{if } i \neq k \tag{6}$$

The result of PCA obtained from the data of typical building in this research is shown in Fig. 6. It is calculated that the first 39 principal components account for more

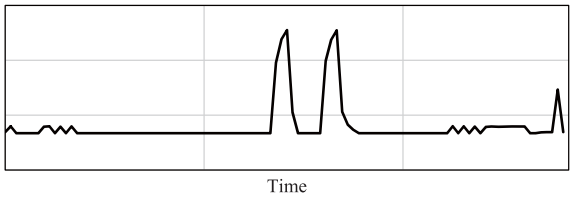

Fig. 5 Time series data

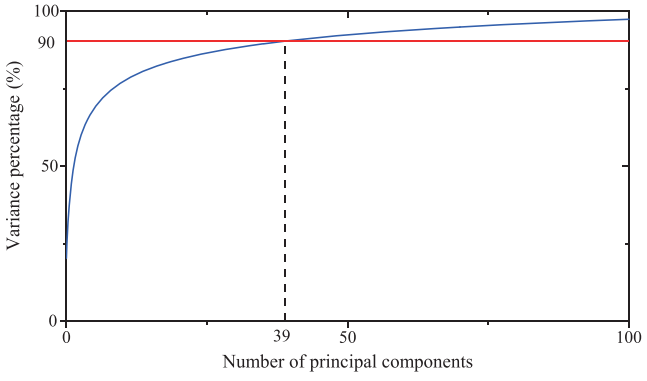

Fig. 6 Variance percentage of the first *i* principal components

than 90% of the total variance value and therefore these components are employed in the following research as well. What should be noted is that the quantity of principal components is also a parameter whose determination should be conducted on the basis of different data sets.

2.5 Classification algorithm

Considering that the input is discrete, XGBoost, a DM library developed by Chen and Carlos (2016), is used for identifying the specific strategies. XGBoost, which integrates a bunch of low accuracy decision tree into a high accuracy model, is developed with the purpose of solving problems related to data science such as ranking, classification and regression and it has been successful applied in many cases (Chen and Carlos 2016). It has been proved that ensemble learning is more efficient and accurate in comparison with traditional algorithm for data mining such as Artificial Neural Network (ANN) while carrying out classification or regression over data about building energy (Foucquier et al. 2013; Chakraborty and Elzarka 2019). As a comparison, the method of Logistic Regression is also applied in the training process of model.

In the algorithm, the trees in the model are trained once a time. The objective function of XGBoost can be shown by the equation as follows:

$$\text{obj} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)}\right) + \sum_{j=1}^{t} \Omega\left(f_i\right) \tag{7}$$

where y_i refers to actual value; $\hat{y}_i$ denotes prediction value; l indicates the loss function; $\Omega\left(f_i\right)$ represents the regularization term; t refers to the prediction step and n denotes the quantity of data points.

The lost function is defined in the equation as follows:

$$l\left(y_i, \hat{y}_i\right) = \sum_i \left(y_i - \hat{y}_i\right)^2 \tag{8}$$

The regularization term is defined in the equation as follows:

$$\Omega\left(f_i\right) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{9}$$

where γ, λ refer to regularization parameters; T denotes the quantity of leaves in the tree; w_j indicates the scores on leaves.

The final solution to the objective function is shown in the equation as follows:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{10}$$

$$\text{obj} = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{11}$$

where H_j, G_j are defined in the equations as follows:

$$H_j = \sum_{i \in I_j} \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{12}$$

$$G_j = \sum_{i \in I_j} \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right) \tag{13}$$

$$I_j = \left\{i \mid q(x_i) = j\right\} \tag{14}$$

3 Case studies

Two case studies are conducted for the purpose of testing the algorithm mentioned above developed by us. One of them is a building in EnergyPlus prototypical and the other is a large commercial building in Shanghai. It is acknowledged that the first step to generate the dataset for testing is implementing necessary control strategy in EnergyPlus properly.

3.1 Implementation of specific control strategy in EnergyPlus model

With the aim of implementing control strategies in EnergyPlus, EMS (Energy Management System) module, which allows developers to build up our own control strategies in a building using the EnergyPlus Runtime Language (Erl), in EnergyPlus is adopted. EMS has two input objects: EnergyManagementSystem: Sensor and EnergyManagementSystem: Actuator. The former object can achieve decision loops in accordance with the outputs of EnergyPlus. For instance, this object can be used by us to decide whether to turn on the HVAC system on the basis of the average indoor temperature in a thermal zone. Then, the latter object can control the HVAC system in accordance with the different values given. For instance, the object can

be employed for the purpose of adapting the temperature of chilled water supply.

3.1.1 Implementation of DR on/off strategy

In this model, the DR period ranges from 13:00 to 15:00. Upon the beginning of DR, all chillers are shut down and therefore indoor air will get warmer after being gaining heat. The setpoint of indoor air temperature is 24 °C and if the indoor temperature in any room goes beyond 28 °C, then the chillers will be started again. It is shown by researches and experiments that if the chillers are shut down just when the air temperature comes to 24 °C, then the air will get heat soon and arrive at 28 °C for the reason that the thermal mass has not been cooled enough to offset the gained heat. Therefore, the chillers should run for at least 1 hour.

To develop the control algorithm with EMS, EnergyManagementSystem: Sensor objects are created for each zone so as to obtain the indoor temperature value and thus the highest temperature is obtained. Then we are required to create EnergyManagementSystem: Actuator objects for each chiller for the purpose of controlling its on/off state. The two objects are documented in "idf" file as follows:

```
EnergyManagementSystem:Sensor,
  ZMA_Core_bottom,          !- Name
  Core_bottom,              !- Output:Variable or Output:Meter Index Key Name
  Zone Mean Air Temperature; !- Output:Variable or Output:Meter Name
```

```
EnergyManagementSystem:Actuator,
  Chil1_Disptch,                          !- Name
  CoolSys1 Operation Scheme:CoolSys1 Chiller1, !- Actuated Component Unique Name
  Plant Equipment Operation               !- Actuated Component Type
  Distributed Load Rate;                  !- Actuated Component Control Type
```

The pseudocode of the algorithm is listed as follows:

```
Calculate the maximum temperature in all HVAC zones T_zone,max
if during DR period
   if chillers are on
      shut off chillers
   elseif chillers are off
      if T_zone,max > 28°C
         restart chillers
      else no operation
      endif
   elseif chillers are on
      if chillers have been on over 1 hour
           shut off chillers
      else no operation
      endif
   endif
else
   apply routine control strategy
endif
```

In view that a total of three chillers are available in the system, the sequence control of chillers is subjected to the cooling demand. In this paper, and the calculation of the cooling demand is conducted by EnergyPlus as shown in Fig. 7. For instance, when the cooling demand is lower than 1758 kW, only Chiller 1 will be activated.

3.1.2 Implementation of DR reset strategy

As what is mentioned above, the regular temperature of chilled water supply is 6.7 °C, and the temperature is reset to 10 °C during the DR period. In addition, when the indoor temperature goes beyond 28 °C, the supply temperature will return to 6.7 °C again. To develop this control algorithm with EMS, EnergyManagementSystem: Sensor objects are created for each zone for the purpose of acquiring the indoor temperature value and then the highest temperature is obtained. Then the EnergyManagementSystem: Actuator objects is also created with the aim of resetting the supply temperature. The algorithm can be realized by replacing the action of shutting off the chillers with the reset procedure.

```
EnergyManagementSystem:Actuator,
  CLGSETP_SCH_Actuator,          !- Name
  CLGSETP_SCH_Yes_Optimum,       !- Actuated Component Unique Name
  Schedule:Compact,              !- Actuated Component Type
  Schedule Value;                !- Actuated Component Control Type
```

3.1.3 Implementation of the reset strategy for chilled water

Different from the control strategy mentioned previously, the reset strategy for chilled water is able to reset the supply temperature by taking into account the outdoor air dry-bulb temperature. Thus, an EnergyManagementSystem:Sensor abject is created to request this value:

```
EnergyManagementSystem:Sensor,
  OAT,                           !- Name
  Environment,                   !- Output:Variable or Output:Meter Index Key Name
  Site Outdoor Air Drybulb Temperature; !- Output:Variable or Output:Meter Name
```

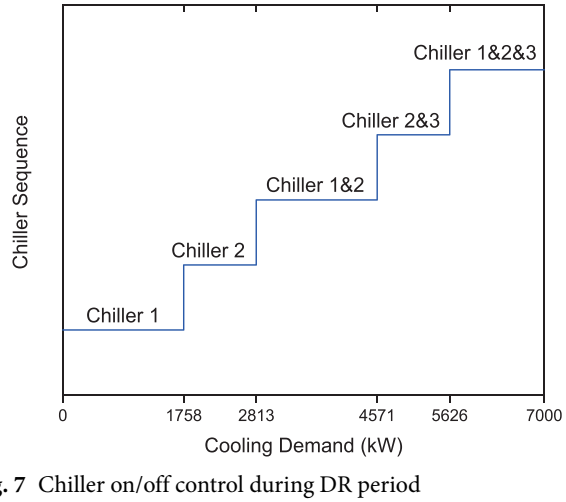

Fig. 7 Chiller on/off control during DR period

3.2 Typical building case

3.2.1 *Building information*

With the purpose of training the identification model, a typical building EnergyPlus model is built up so as to obtain the training data. The model of building which is a typical 13-story (one-story basement included) office building with the area of 3567 m^2 for each floor is established based on DOE commercial reference building program—Large Office (Climate Zone 1A Miami, Florida) (DOE 2012). With the windows evenly distributed, the WWR (window to wall ratio) is 40% on four erect walls. Besides, each story of the building is split into an inner zone and an outer zone by inner walls, and the material walls are set in accordance with ASHRAE 90.1.

As is shown in Fig. 8, the typical HVAC water system whose specific control strategies can be identified based on the identification framework proposed is composed of several chilled water pumps of chillers, condenser water pumps, cooling pumps and AHUs.

3.2.2 *Results obtained from the implementation of control strategies*

With the previously mentioned EnergyPlus model and strategy implementation algorithm as the basis, three different control algorithms are implemented (on/off control of chiller during DR period, chilled water reset during DR period and ASHRAE 90.1 chilled water reset). Figure 9 shows the supply temperature of the typical building simulated after the application of the three control strategies. As shown in Fig. 9(a)), when the chillers are shut off during DR period, the temperature of chilled water supply increases quickly and arrives at nearly 25 °C upon the completion of DR. Figure 9(b) shows that the temperature of chilled supply reaches 10 °C during the DR period. Figure 9(c) shows that the supply temperature changes along with the passage of time due to the change of outdoor temperature. It is easy for us to come to the conclusion from the pictures that all the three control strategies are well performed on the EnergyPlus platform.

It is widely acknowledged that most commercial buildings either use sequential control or coordinated control to satisfy specific requirement of control. The first step carried out by us is developing an algorithm that can be employed to identify these specific control strategies.

3.3 Case study in a large commercial building

The building model is a 31-floor office building (one basement included) which is located in Shanghai (cooling dominated), with a height of 140 m and an area of about 2500 m^2 for each floor. Apart from that, each floor has a room for HVAC plants and the 15th floor is employed for the establishment of install equipment such as plate heat exchangers, pumps, etc.

3.3.1 *HVAC system*

The entire building is equipped with a chilled water system including two 2813 kW centrifugal chillers and a 1758 kW centrifugal chiller. Similar to Fig. 2, the water loops, which consist of typical primary/secondary pumping systems, are illustrated in Fig. 10. In addition, the temperature of chilled water supply is constant (6 °C). The information about the chilled water system is listed in Table 2 in detail. The condenser water loop is composed of three cooling towers and three water pumps of constant condenser. Moreover, two AHUs are set in every story so as to supply air to the VAV Boxes in two thermal zones (inner zone and outer zone).

3.3.2 *Model calibration*

In the process of calibration, which was carried out to ensure that the building model is able to simulate the building, a simple control strategy (thermostat on/off control), instead of specific control strategies, was applied. Figure 11 shows both the measured data and the practical data of lighting

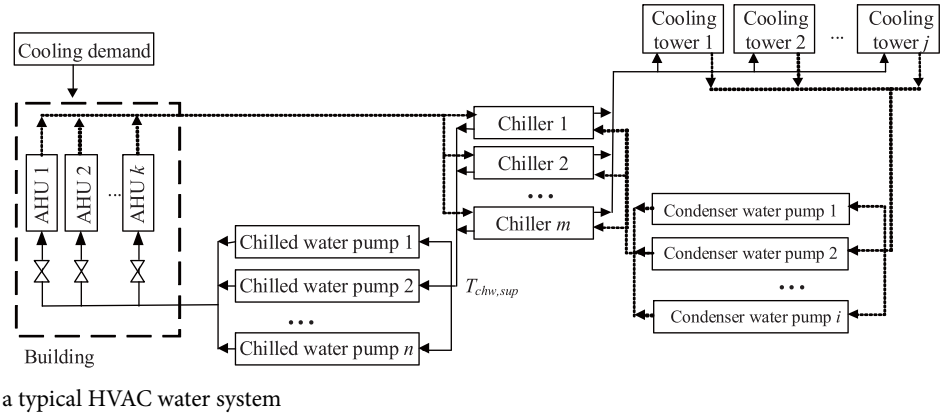

Fig. 8 Diagram of a typical HVAC water system

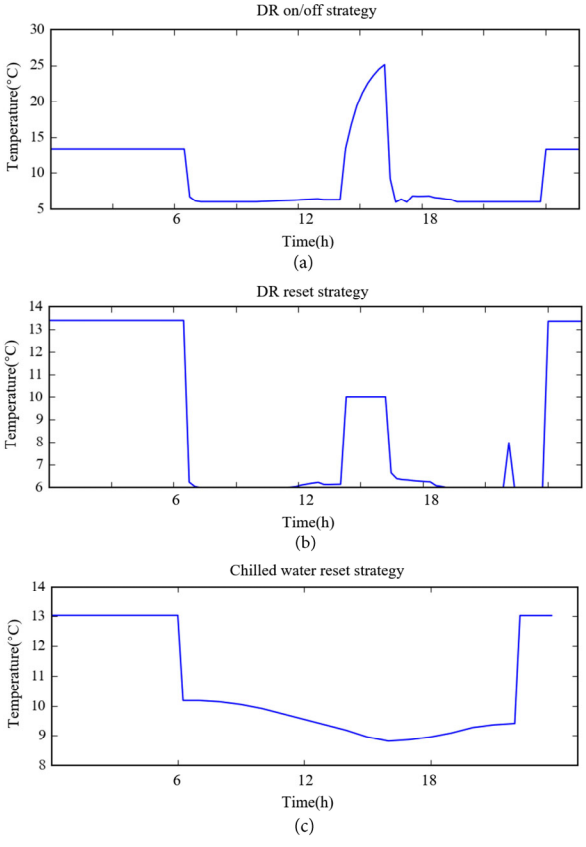

Fig. 9 Temperature of chilled water supply under three control strategies

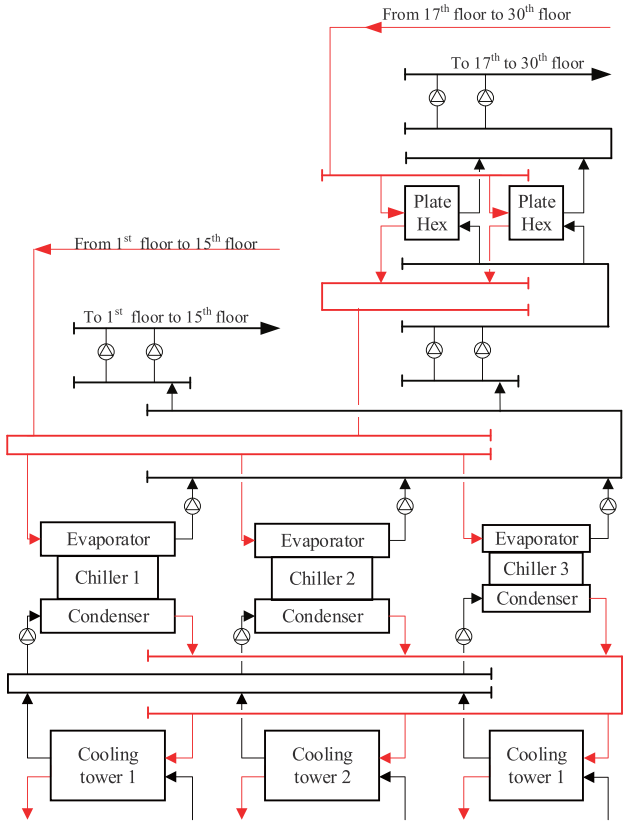

Fig. 10 HVAC system of the office building

Table 2 Parameters of the chilled water system

Chillers	Equipment	Number	Cooling capacity (kW)		Power (kW)
	Centrifugal chiller	1	1758		314
	Centrifugal chiller	2	2813		504
Cooling towers	Equipment	Number	Flow (m^3/h)		Power(kW)
	Open loop cooling tower	2	569		7.5×4
	Open loop cooling tower	1	392		5.5×3
Pumps	Equipment	Number	Flow (m^3/h)	Head (m)	Power (kW)
	Primary pump	2	403	19	30
	Primary pump	1	252	15	15
	Secondary pump (for high floor)	2	351	26	37
	Secondary pump (for low floor)	2	280	28	37
	Chilled water pump (plate HEx)	2	278.5	33	45
	Water pump condenser	2	581	34	75
	Water pump of condenser	1	366	33	55

and equipment power and it is shown that the Mean Absolute Percentage Error (MAPE) is lower than 5%. Since there is some inaccurate input such as occupancy data and equipment usage rate, the result obtained from simulation is acceptable and therefore the model is reliable.

Figure 12 shows both the simulated and measured input power of the chillers in July and August, respectively, and it can be observed that the two lines are very close to each other and the Mean Absolute Percentage Error (MAPE) is lower than 5%. In this process of evaluation, the annual

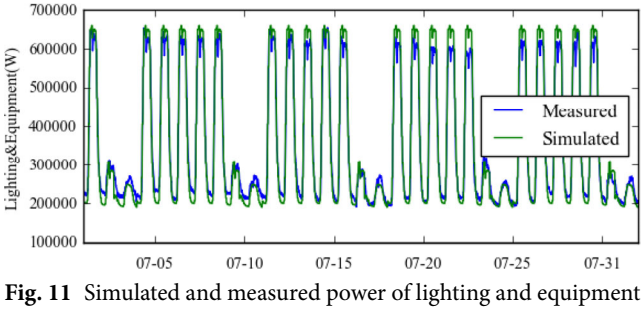

Fig. 11 Simulated and measured power of lighting and equipment

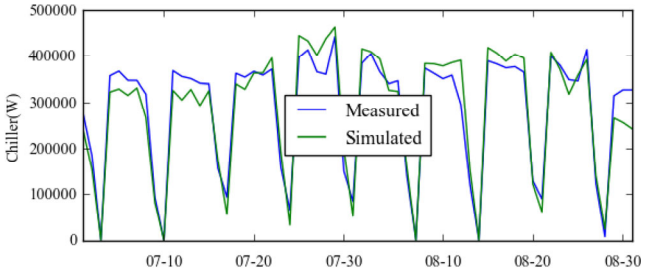

Fig. 12 Simulated and measured power of chillers

hourly on/off states of each chiller are obtained on the basis of the electricity data. Then EnergyPlus can implement this control strategy after reading the state file.

3.3.3 *Data generation*

After the calibration, the method of applying three different specific control strategies which were discussed in Section 3.1 was applied for the aim of generating the raw data of the case study. After that, both the identical feature extraction and PCA transformation were employed so as to satisfy the input requirement of the model.

4 Results obtained from identification of control strategy

4.1 Results of cross validation

Upon completing the preprocessing, about 360 pieces of data related to time series can be generated in a cooling season for a single building. With the purpose of realizing more reasonable evaluation of the accuracy of XGBoost algorithm, *k*-fold cross validation is applied. The data set employed to evaluate the model is then split into a total of *k* groups randomly and each of the *k* groups in the following *k* validation serves as a test data set and the rest *k*−1 groups function as a training set together. In this way, an overall evaluation of the model can be obtained from the given data set.

The evaluation over the identification accuracy of the models is shown in Eq. (15) as follows:

$$\text{Accuracy} = \frac{n_{\text{correct}}}{n_{\text{sample}}} \tag{15}$$

where n_{sample} refers to the quantity of samples, and n_{correct} denotes the quantity of the samples that are correctly identified in the model.

The results obtained from two identification models are shown in Fig. 13. It is shown from the boxplot that the accuracy of both Logistic Regression model and XGBoost model is subjected to great fluctuation and the mean accuracy obtained from the two models are approximately 90% and

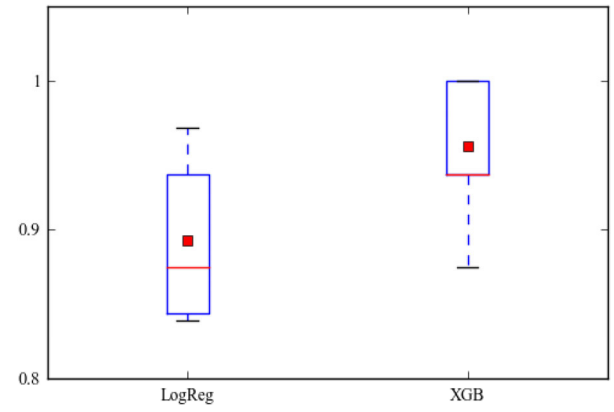

Fig. 13 Accuracy of *k*-fold cross validation

95%, respectively. Since XGBoost shows more reliable performance, this model was employed to test in our practical case.

4.2 Practical case identification results

Figure 14 shows the result obtained from identification of the large commercial building model in which cross marks indicate the predicted values and circle marks represent the actual values. In addition, the predicted value is deemed as wrong if no cross mark is located in the center of a circle mark. As mentioned above, strategy 1, 2, 3 and 0 refer to DR on/off strategy, DR reset strategy, chilled water temperature reset strategy, and no specific strategy, respectively. In 40 simulated data sets including strategies 1, 2, 3, 0, 3 wrong predicted values are obtained and therefore the identification accuracy obtained is 92.5% in accordance with Eq. (15).

Considering that the control strategy in adjacent days is usually fixed, the mode of identification result of several adjacent days can be employed for the purpose of improving the accuracy. Therefore, a total of 11 groups of data sets are generated in the process, and 3 adjacent days' data is available for each group which is applied to the same strategy. As shown in Fig. 15, blue lines indicate the results obtained from identification of these 11 datasets. The principle for 3 adjacent days' identification is that if more than 2 days' data in 3 adjacent days is identified as being obtained by

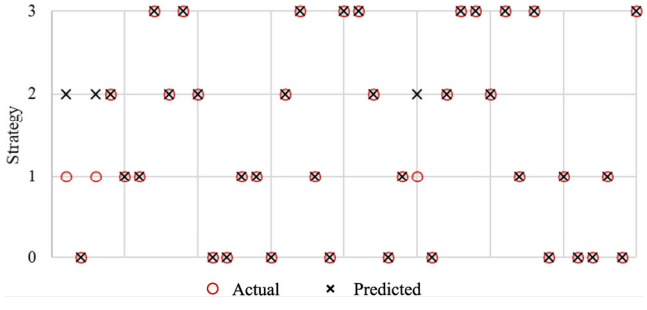

Fig. 14 Results from identification using one-day data

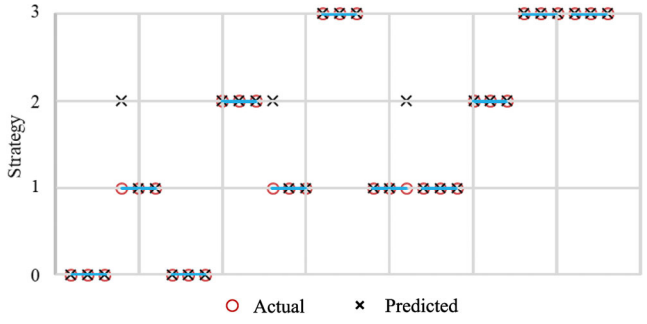

Fig. 15 Results obtained from identification using 3 adjacent days' data

using a specific strategy, then the strategy is deemed as the strategy adopted in the 3 days. Taking the second blue line from the left of the figure as an example, it is easy to observe that the strategy of the first day is incorrectly identified as strategy 2 and the result obtained from the rest two days is correct. In accordance with the principle, it is then known that the final predicted strategy is strategy 1. Despite the fact that some single days are not identified correctly in this test, the algorithm can achieve an accuracy of 100% in these 11 datasets with the data from 3 adjacent days.

5 Conclusion

In this research, some DM algorithms are proposed for the purpose of identifying the control strategies in large commercial buildings. These identified strategies can serve as the basis in analyzing and improving building operation.

The major contribution made by this research can be summarized as follows:

(i) A framework was developed to identify the specific control strategies on the basis of data mining algorithm on system level. In this framework, a typical building model with HVAC control strategies and three specific control strategies was built up and then it was applied to the model through Erl. The results obtained from simulation could well reflect the effects of different strategies. Then two algorithms were applied to preprocessing the raw data generated from EnergyPlus and therefore the feature of the raw data could be well preserved and the dimension could be reduced. Finally, both the XGBoost and Logistic Regression were employed to train the model.

(ii) It is shown by the *k*-fold cross validation that the identification method employed is able to achieve an accuracy of 95% using XGBoost algorithm, which is obviously higher than that of Logistic Regression (90%). With the purpose of testing the identification algorithm, the case model was developed and calibrated on the basis of submetering data, then the test data was obtained accordingly through the same procedure of simulation and preprocessing. It is then shown by the result that the precision of identification can reach 92.5% if the data collected is used in a day and 100% if 3-day data is employed.

In this research, the identification method is tested by adopting the data generated from the EnergyPlus which has been calibrated from the practical building. In the future, the framework will be further tested in real buildings by employing the measured data of different specific control strategies, and the feasibility about whether the present framework can be further extended to buildings of other kinds except for office buildings. Furthermore, specific control strategies may also be added to the framework as well.

Acknowledgements

This work was supported by the National Science & Technology Pillar Program during the thirteenth Five-year Plan Period (No. 2017YFB0903404).

References

ASHRAE (2011). Handbook of Fundamentals. Atlanta, GA, USA: American Society of Heating, Refrigeration, and Air-Conditioning Engineers.

ASHRAE (2013). Energy Standard for Buildings Except Low-Rise Residential Buildings. ANSI/ASHRAE/IES Standard 90.1-2013. Atlanta, GA, USA: American Society of Heating, Refrigeration, and Air-Conditioning Engineers.

Bauer M, Scartezzini JL (1998). A simplified correlation method accounting for heating and cooling loads in energy-efficient buildings. *Energy and Buildings*, 27: 147–154.

Braun JE (1989). Applications of optimal control of chilled water systems without storage. *ASHRAE Transactions*, 95(1): 663–675.

Chakraborty D, Elzarka H (2019). Early detection of faults in HVAC systems using an XGBoost model with a dynamic threshold. *Energy and Buildings*, 185: 326–344.

Chen T, Carlos G (2016). XGBoost: A scalable tree boosting system. Paper presented at the 22nd ACM SIGKDD International Conference, San Francisco, USA.

CIBSE (2000). Building Control Systems. London: Routledge.

D'Oca S, Hong T (2014). A data-mining approach to discover patterns of window opening and closing behavior in offices. *Building and Environment*, 82: 726–739.

D'Oca S, Hong T (2015). Occupancy schedules learning process through a data mining framework. *Energy and Buildings*, 88: 395–408.

Dony RD, Wesolkowski S (1999). Edge detection on color images using RGB vector angles. In: Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering, Edmonton, Canada, pp.687–692.

DOE (2012). Commercial Reference Buildings. US Department of Energy. Available at https://www.energy.gov/eere/buildings/commercial-reference-buildings.

EIA (2016). International Energy Outlook 2016. Available at https://www.eia.gov/outlooks/archive/ieo16. Accessed May 11 2016.

EIA (2018). Commercial Buildings Energy Consumption Survey. Available at http://www.eia.doe.gov/emeu/cbecs/contents.html.

Fan C, Xiao F, Yan C (2015). A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Automation in Construction*, 50: 81–90.

Feng F, Li Z (2017). A methodology to identify multiple equipment coordinated control with power metering system. *Energy Procedia*, 105: 2499–2505.

Foucquier A, Robert S, Suard F, Stéphan L, Jay A (2013). State of the art in building modelling and energy performances prediction: a review. *Renewable and Sustainable Energy Reviews*, 23: 272–288.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457: 1012–1014.

Hong WC (2009). Electric load forecasting by support vector model. *Applied Mathematical Modelling*, 33: 2444–2454.

Kalogirou S, Neocleousp C, Schizas C (1997). Building heating load estimation using artificial neural networks. In: Proceedings of the International Conference CLIMA 2000.

Kanopoulos N, Vasanthavada N, Baker RL (1988). Design of an image edge detection filter using the Sobel operator. *IEEE Journal of Solid-State Circuits*, 23: 358–367.

Katipamula S, Brambley M (2005). Review article: methods for fault detection, diagnostics, and prognostics for building systems—A review, part II. *HVAC&R Research*, 11: 169–187.

Kusiak A, Li M, Zhang Z (2010). A data-driven approach for steam load prediction in buildings. *Applied Energy*, 87: 925–933.

Li Q, Meng Q, Cai J, Yoshino H, Mochida A (2009). Applying support vector machine to predict hourly cooling load in the building. *Applied Energy*, 86: 2249–2256.

Li M, Miao L, Shi J (2014). Analyzing heating equipment's operations based on measured data. *Energy and Buildings*, 82: 47–56.

Li W, Xu P, Lu X, Wang H, Pang Z (2016). Electricity demand response in China: Status, feasible market schemes and pilots. *Energy*, 114: 981–994.

Li G, Hu Y, Chen H, Li H, Hu M, Guo Y, Liu J, Sun S, Sun M (2017). Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions. *Applied Energy*, 185: 846–861.

Mirzaei A, Reza S (2012). A data mining framework for extracting product sales patterns in retail store transactions using association rules: A case study. *Journal of American Science*, 8(9): 304–308.

Morar A, Moldoveanu F, Groller E (2012). Image segmentation based on active contours without edges. In: Proceedings of the Intelligent Computer Communication and Processing (ICCP2012), Cluj-Napoca, Romania, pp. 213–220.

Motta Cabrera DF, Zareipour H (2013). Data association mining for identifying lighting energy waste patterns in educational institutes. *Energy and Buildings*, 62: 210–216.

Niu D, Wang Y, Wu DD (2010). Power load forecasting using support vector machine and ant colony optimization. *Expert Systems With Applications*, 37: 2531–2539.

Pal NR, Pal SK (1993). A review on image segmentation techniques. *Pattern Recognition*, 26: 1277–1294.

Qiu S, Feng F, Li Z, Yang G, Xu P, Li Z (2019). Data mining based framework to identify rule based operation strategies for buildings with power metering system. *Building Simulation*, 12: 195–205.

Rafael CG, Richard E, Woods, Steven (2009). Digital Image Processing Using MATLAB®. Knoxville, TN, USA: Gatesmark Publishing.

Wang D, Zhou S (2008). Color image recognition method based on the Prewitt Operator. In: Proceedings of the International Conference on Computer Science & Software Engineering, Colombo, Sri Lanka.

Wang H, Lu X, Xu P, Yuan D (2015). Short-term prediction of power consumption for large-scale public buildings based on regression algorithm. *Procedia Engineering*, 121: 1318–1325.

Witten IH, Frank E (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Burlington, MA, USA: Morgan Kaufmann Publishers.

Yu Z, Haghighat F, Fung BCM, Yoshino H (2010). A decision tree method for building energy demand modeling. *Energy and Buildings*, 42: 1637–1646.

Yu Z, Haghighat F, Fung BCM, Zhou L (2012). A novel methodology for knowledge discovery through mining associations between building operational data. *Energy and Buildings*, 47: 430–440.