BUILD SIMUL
https://doi.org/10.1007/s12273-018-0472-6

Data mining based framework to identify rule based operation strategies for buildings with power metering system

Shunian Qiu, Fan Feng, Zhengwei Li (✉), Guang Yang, Peng Xu, Zhenhai Li

School of Mechanical Engineering, Tongji University, Shanghai, China

Abstract

Operation strategies influence the building energy efficiency. In order to enhance the building energy efficiency, it's necessary to adopt proper operation strategies on building equipment. Thus, the identification of existing operation strategies is necessary for the improvement of operation strategies. A data mining (DM) based framework is proposed in this paper to automatically identify the building operation strategies. The framework includes classification and regression tree (CART), and weighted association rule mining (WARM) method, targeting at three types of rule based control strategies: on/off control, sequencing control (for equipment of the same type), and coordinated control (for equipment of different types). The performance of this framework is validated with power metering system data and manual identification results based on on-site survey of three buildings in Shanghai. The validation results suggest that the proposed framework is capable of identifying building operation strategies accurately and automatically. Implemented on the original software named BOSA (Building Operation Strategy Analysis), this framework is promising to be used in engineering field to enhance the efficiency of building operation strategy identification work.

Keywords

power metering system,
operation strategy,
data mining,
automatic identification,
weighted association rule mining,
classification and regression tree

Article History

Received: 20 January 2018
Revised: 10 August 2018
Accepted: 13 August 2018

1 Introduction

Building sector consumes more than 30% of the total energy worldwide (EIA 1989). An efficient way to alleviate global warming and improve environmental sustainability is to enhance building energy efficiency. In order to enhance the building energy efficiency, it's important to analyze the energy consumption and find out the operation problems of the building.

The operation strategies of the building equipment could significantly influence the energy consumption of the building. For instance, Beghi et al. (2011) applied a multi-phase genetic algorithm to optimize the operation of the multiple-chiller system, which effectively reduced the energy consumption of the HVAC system. Thus, in order to reduce building energy consumption, it's necessary to apply proper operational strategies to building appliances. In order to accomplish that, the strategies of existing building appliances should be identified so that the inappropriate strategies could be replaced.

In buildings, operation strategies typically include two types: rule based strategy and model based strategy. Although model based strategy may lead to higher performance, its use is still limited due to reliability and robustness issues (Sun et al. 2013).

In order to identify the operation strategies in a building, proper input data is required. Power submetering system, which is an effective way of monitoring and supervising the energy consumption in buildings (O'Driscoll and O'Donnell 2013), provides the energy consumption data valuable for analysis. A typical power submetering system measures the electricity consumption of four major building service systems: lighting and plug loads; heating, ventilation, and air conditioning (HVAC) systems; power system (fan, pump, etc.); and special equipment (such as kitchen, information center, etc.). Compared with building automation system (BAS) system, this system is more reliable and easy to maintain. Also, the submetering data is easy to analyze. Due to the advantages of power submetering system, it has been adopted as a recommended energy monitoring system in

E-mail: zhengwei_li@tongji.edu.cn

List of symbols

X	input variables	d	distance between output variables and prediction variables
Y	output variables	k	number of regions divided by a tree
R	a particular region where the input variables lie	l	tree level
W	pooled within-group sum of squares	μ	average of samples
G	gap statistics	σ	standard deviation of samples
z	normalized score		

many countries. Taking China as an example, according to the survey data by Chinese Construction Industry Association, by Dec 2014 the number of public buildings equipped with power metering systems has exceeded 7400 (CCIA 2015).

Due to the significant data size of submetering system, manual analysis is no longer a feasible approach. Thus, computer aided techniques have to be resorted. Data mining (DM) is a powerful approach to deal with "big data", and has been successfully applied to understand human preferences (Witten and Frank 1999). In the area of building energy research, DM methods have been heavily used for energy demand prediction (Ku and Jeong 2018; Li and Huang 2013; Magoulès et al. 2013). Also, DM methods are applied to the analysis of equipment operation strategies and occupants' behavior (Dong et al. 2018).

Yu et al. (2012) firstly adopt association rule mining (ARM) to examine the relationship between building operation data. In their research around 500 if-then rules were obtained, which were then filtered manually to find the problematic behavior. Obviously, manually filtering a set of candidate rules is not practical when the number of rule sets is large. To reduce the number of rule candidates, Fan et al. (2015) proposed to conduct an energy use profile based clustering as a preliminary step, so that ARM is performed only for a representative data set. Motta Cabreba and Zareipour (2013) implemented a pattern recognition algorithm to identify abnormal light use behavior in classrooms, and found a high correlation between high energy consumption and light on during unoccupied period. D'oca and Hong (2015) adopted decision tree, rule induction algorithm and cluster analysis to obtain the consistent pattern of occupancy schedules which could be applied to the building energy simulation. Pang et al. (2018) adopted Fourier Series Regression to identify the occupancy schedule of a commercial building with the mobile positioning occupancy data, which could be used in building energy simulation to enhance the model accuracy. Li et al. (2014) used 4 data mining methods to identify the operation strategy of heating equipment. The identified strategy was used to predict the cycle time and idle time of the equipment in the future.

Since the building operation strategies consist of human behaviors (e.g., light on-off control behaviors), and equipment operation strategies (e.g., chiller operation strategies), it can be seen that DM methods are promising to be used to analyze building operation.

This paper proposes a DM based framework to identify rule based building operational strategies with power submetering system data. The proposed framework is intended to automatically identify existing building operation strategies so that the improper strategies could be found and replaced. In doing so, building energy efficiency could be improved. Implemented on an original software BOSA (Building Operation Strategy Analysis), this framework is promising to be applied in engineering field to enhance the identification efficiency and free manual labor of engineers.

The content of this paper is organized as following: first, typical building operational strategies are defined and classified into three different types (on/off control, sequencing control, and coordinated control); second, a systemic DM based framework targeting at these operation strategies is proposed; third, the software to implement this framework, BOSA, is introduced; then, case studies are conducted to validate the effectiveness of the proposed approach; finally, the research is concluded.

2 Typical building operational strategies

In general, building operational strategies can be classified into three groups: individual equipment control, multiple equipment sequencing control, and multiple equipment coordinated control (as shown in Fig. 1).

More specifically, individual equipment control strategies can be divided into two categories: on-off control and continuous control (also termed ratio control in this study). While the former is often based on occupancy presence, the latter is typically triggered by certain parameters (such as frequency, speed, or static pressure). Compared with on/off control, ratio control is a more advanced control paradigm, requiring modern building automation system (BAS) as the supporting backbone, which means the operation of ratio

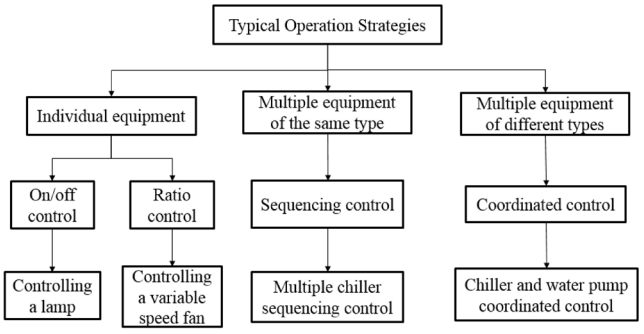

Fig. 1 Typical building operational strategies

controlled equipment is highly related to the BAS signal and data. Since the work of this paper is based on power submetering system data, the identification of this type of control strategy (ratio control strategy) is not discussed in this research.

Strategies to control equipment of the same type are referred to as sequencing control. This type of strategy is mostly seen in primary plants of air conditioning systems. For example, when the cooling load is smaller than the full capacity of the chillers, a decision has to be made regarding how to distribute the load to chillers (Beghi et al. 2011). Since the performance of cooling source depends strongly on the part load ratio, this type of strategy could significantly affect the energy consumption.

Strategies to control equipment of different types are referred to as coordinated control in this study. This type of strategy is typically set to guarantee the safety of equipment. A typical example is the coordination of chillers and pumps, which is intended to avoid chillers' malfunction caused by water flow mismatch.

All of these three strategies can be executed either based on rules or more advanced models with the aim to optimize certain variables. The DM based framework to identify the rule based strategies is demonstrated in the next section.

3 Proposed identification framework

The proposed framework is illustrated in Fig. 2, which consists of three main steps: pre-processing (to detect and diagnose the data quality of the power metering data), strategy identification and post-processing.

3.1 Input and output data set

As mentioned in Section 2, in this research, building operational strategies are divided into three categories: individual equipment control, sequence control, and coordinated control. Some of them are triggered by time (e.g., scheduling control), some of them are triggered by environment parameters (e.g., cooling tower fan frequency

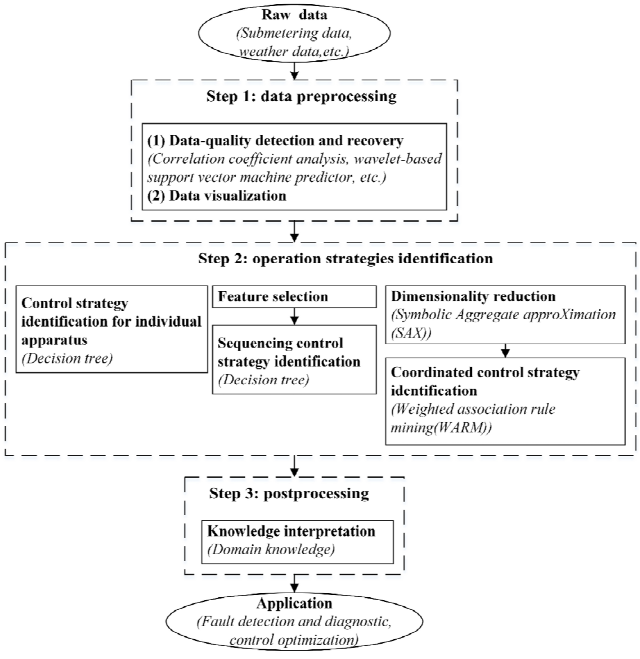

Fig. 2 Framework for rule based strategy identification

control based on outdoor wet bulb temperature). According to the trigger variable type of each kind of strategies, the input variables required to identify the operation strategies are categorized and listed in Table 1.

3.2 Data pre-processing

In the preprocessing step, three types of faults are detected and diagnosed: (1) NAN (or zero fault), caused by hardware faults (including faults meter fault, data collection system failure, or transmission system failure); (2) negative fault, where negative data is recorded, caused by an inverse current flow due to improper wiring; (3) disturbance fault, mostly due to environmental disturbance around working meters (Fu et al. 2016).

Among these three types of faults, type 1 fault can be detected and corrected by replacing NAN values with the average of near data points. Type 2 fault could be handled by replacing negative values with the average of near data points. To detect and diagnose the type 3 fault, the box plot based on statistics could be applied. With box plot, the extreme outliers of the data set could be identified and corrected. Similarly, the faulty data points (outliers) are replaced by the average of near data points.

3.3 Strategy identification

As introduction above, this study aims at three types of rule based control strategies: on/off control, sequencing control, and coordinated control. Among typical DM methods, DT

Table 1 Input and output variables for strategy identification

Strategy	Time triggered	Environment triggered
Individual equipment control; sequencing control	Month of year, week of month, day of week, hour of day	Environment parameters (outdoor dry bulb temperature, outdoor wet bulb temperature, etc.) System operation parameters (chiller cooling capacity, etc.)
Coordinated control	Power metering data	Power metering data

has advantages on the visualization of trained model. With the visual of the trained tree, it's easier for analysts to analyze the building operation strategy and explain the energy consumption. The coordinated control strategy is typically revealed by the association of the energy consumption of coordinated appliances, which means this type of operation strategy is suitable for ARM to identify since ARM is intended to recognize the association between several variable (Zhang and Zhang 2002). Thus, DT is selected to identify the on-off control and sequence control, and ARM is selected to identify the coordinated control strategies.

3.3.1 *Decision tree*

In our study, the classical classification and regression tree (CART) algorithm is used for tree growing (Lewis 2000). And Gap statistics is adopted to determine the optimal tree level, thus, overfitting problem could be prevented by pruning (Tibshirani et al. 2001). The method of Gap statistics is demonstrated below. Suppose there are n independent samples with p independent variables and a target variable Y, the distance d between certain samples in terms of Y can then be described with Eq. (1):

$$d = \sum_{j=1}^{n} (Y_j - \overline{Y})^2 \tag{1}$$

Further suppose the original data set is divided into k groups (i.e. k leaves) by the tree, and the number of the data sets in the i-th group is n_i, then the total within distance of these k groups can be calculated by Eq. (2):

$$W = \sum_{i=1}^{k} \frac{d_i}{2n_i} \tag{2}$$

The gap statistic is then computed by comparing W with the expectation under m sample datasets drawn from reference distributions (Tibshirani et al. 2001). In this study, m is chosen as 20. The reference distribution is typically uniform distribution, whose minimum and maximum values are set to the same as in the raw data set. Therefore, the gap statistic is defined as the difference between the W and the expectation.

$$G = \frac{1}{m} \sum_{i=1}^{m} (\log W_{\text{ref}}) - \log W \tag{3}$$

For each tree level, a unique gap value can be calculated, and the optimal tree level l_{op} can then be found based on Eq. (4):

$$l_{\text{op}} = \arg\min_i \{i \mid G_i \geq G_{i+1} - \sigma(G_{i+1})\} \tag{4}$$

where $\sigma(G_{i+1})$ is the standard deviation calculated using m sample datasets.

3.3.2 *Weighted association rule mining*

Association rule mining (ARM) is a DM technique applied mainly to identify hidden relationships in the form of rules (Zhang and Zhang 2002). Two parameters are critically important to identify the potential rules: support and confidence. While support is the joint probability of the antecedent and consequent, and confidence is the conditional probability of consequent, given the antecedent. Typically, rules associated with high support and confidence values are considered as important rules, which should be given first consideration when checking the rule validity.

However, it is found that when the volume of a dataset exceeds a certain amount, the speed of the commonly used Apriori algorithm will decrease significantly. In our study, we use a Lenovo computer with 12 GB RAM memory and Intel i5 processor to deal with the data set (8760 observations with 17 variables). We get more than 10000 rules and an "out of memory" error. To speed up the process, the original dataset should be condensed to a more efficient form. In this paper, Symbolic Aggregate approXimation (SAX) is adopted to perform the dataset condensation (Miller et al. 2015). Besides, using SAX helps us to capture the daily profile with reduced number of identified rules, and thus helps to reduce the work load in further interpretation.

The SAX technique is composed of several steps, as illustrated in Fig. 3. First, the original hourly data stream is normalized with Z-score normalization (as formulated in Eq. (5), μ is the mean of this data stream, σ is the standard deviation, $Z(t)$ is the normalized data).

$$Z(t) = \frac{x(t) - \mu}{\sigma} \tag{5}$$

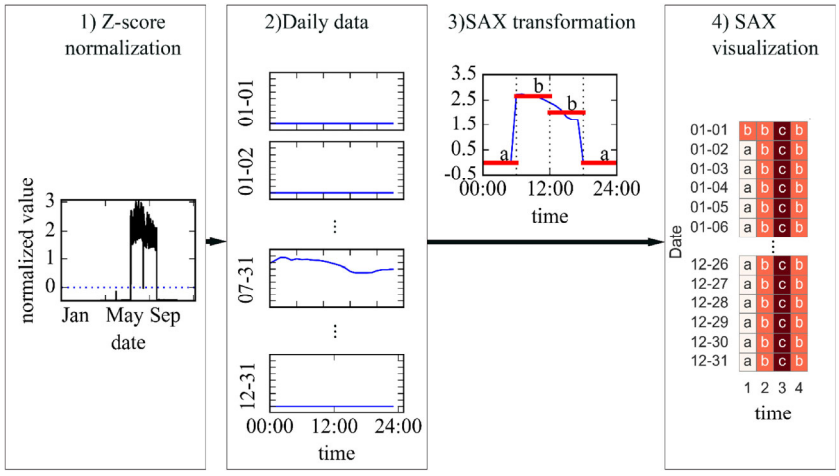

Fig. 3 Diagram for SAX processing

Second, the hourly normalized data are broken down into n individual non-overlapping subsequences, and each subsequence consists of 24 data points, which are hourly energy consumption data in a day. Each subsequence is further divided into W equally sized segments. The average of the data across each segment is calculated and an alphabetic character is assigned to each of these segments according to their data averages. For example, character "a" means the lowest level, character "b" means a higher level than "a", and so on. Here, another parameter A is introduced to denote the number of discretization levels. It should be noted that, for a discrete variable, the number of discretization levels is equal to the number of all possible occurrences.

For a dataset with p variables, the processing above needs to be repeated for each variable. After concatenating the strings of these variables horizontally, 365 sentences can be derived, each of which consists of p words (the number of variables). These 365 sentences can be clustered into several groups, each of which represents a typical day and is given a weight to indicate its frequency. This process is illustrated in Fig. 4.

For the derived weighted dataset, conventional ARM is not suitable, as it gives each instance equal weight. To solve this problem, Weighted Association Rule Mining (WARM) is deployed (Feng 2003). Different from conventional ARM, support in WARM is the ratio of the weight of the observations that contain both antecedent and consequent to the weight of all observations. Similarly, confidence is the ratio of the weight of the observations that contain both antecedent and consequent to the weight of observations that only contain antecedent.

3.4 Knowledge interpretation

Once the three types of control strategies (on/off, sequencing, and coordinated control) are identified, the next step is to evaluate the reasonability and effectiveness of these strategies. For instance, if the chiller on/off control is found to be strongly dependent on ambient temperature, then the threshold temperature needs to be further evaluated. This enables further opportunities for energy saving and fault detection.

It should be noted that, the correctness of the identified strategy depends on if all input parameters are included. In case an important input parameter is omitted during the analysis, the identified strategy could be totally different from the real operational strategy. To solve this problem, it is

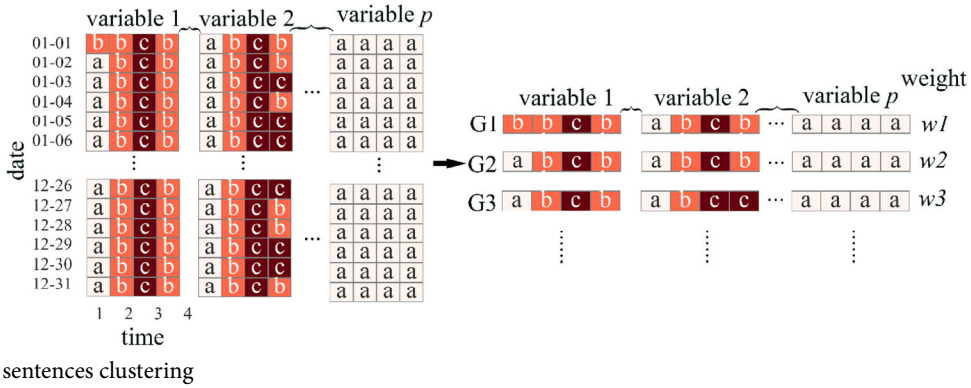

Fig. 4 Daily SAX sentences clustering

suggested to verify the results with building facility managers before implementing any correction measures.

3.5 Building Operational Strategy Analysis tool (BOSA)

BOSA is an Excel based tool that has been developed based on the algorithm mentioned above, whose interface is shown in Fig. 5.

As shown in Fig. 5, BOSA consists of three main parts: data preparation, data preprocessing, and data analyzing. The first step is to prepare input data. At this stage, users enter required input information (such as electricity submetering data, time, environment parameter, etc.). Then, at the preprocessing stage, BOSA detects the quality of the data, and visualize the data to identify potential problems. Finally, to analyze the data, the type of operational strategy to be analyzed needs to be selected. After the computation is done, results show up in a new Excel sheet, as shown in Fig. 6.

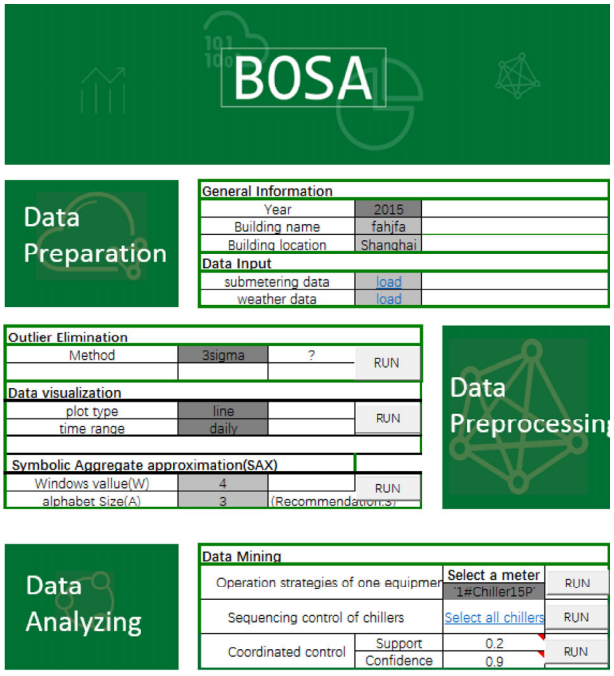

Fig. 5 Main interface of BOSA

	number	precedent		concedent	support	confidence
2	1	2#chiller =" high"	-->	chilledPump2 = " high"	0.2	0.949
3	2	chilledPump2 = " high"	-->	2#chiller =" high"	0.2	1
4	3	2#chiller =" low"	-->	chilledPump2 = " low"	0.8	0.99
5	4	chilledPump2 = " low"	-->	2#chiller =" low"	0.8	1
6	5	2#chiller =" low"	-->	coolingPump1 = " low"	0.8	0.999
7	6	coolingPump1 = " low"	-->	2#chiller =" low"	0.8	0.947

Fig. 6 An example of coordinated control strategy identification results

4 Case study

In this section, three case studies are presented to illustrate how the proposed framework can be applied to real power metering data, each aims at a specific type of operational strategy. Also, the identification result of each case study is validated by measured power metering system data, appliance operator and experienced expert who manually identified the operation strategies based on on-site survey information.

4.1 Case study 1: lighting system on/off control

The first case is an office building located in Changning District, Shanghai city. For each floor of the building, the electricity consumption of lighting, air conditioning, and other end uses is metered. In this study, the electricity consumption of lighting in one of the floors in 2015 is analyzed.

The first step is to detect the data quality of the monitoring data. From Fig. 7 (left), it can be seen that there are some extreme outliers in the original data set of hourly lighting energy consumption. Using the data quality FDD methods introduced in Section 3.2, those outliers are replaced with the average values of near data point values. The result is shown in Fig. 7 (right).

The second step is to develop an optimal tree for the lighting system operational data. Time (e.g., 9:00, 15:00–19:00), month (e.g., March, April), date (e.g., 20th), and day type (e.g., weekday or weekend/holiday) are taken as input variables. The lighting system energy consumption data measured by power metering system is taken as output variables. Thus, data points like {"Time": 9:00, "month": March, "date": 20th, "energy": 15 kWh} are used to train the DT. As is shown in Fig. 8, when the tree level exceeds 13, the lower bound of the Gap statistic value falls below the Gap statistics with tree level at 13. Therefore, the optimal tree level is determined to be 13. Pruned by Gap statistics, the optimal DT is shown in Fig. 9.

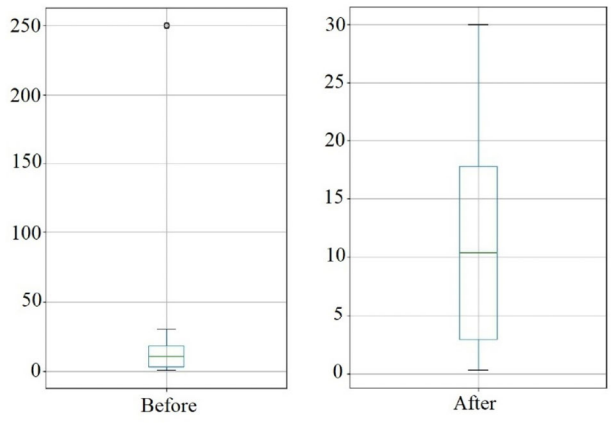

Fig. 7 Annual power metering data of the lighting system (left: original data set; right: pre-processed data set; data value is the hourly energy consumption (kWh))

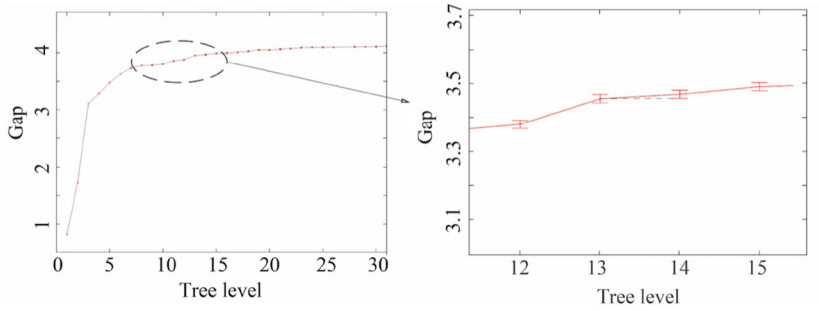

Fig. 8 Searching for an optimal tree structure

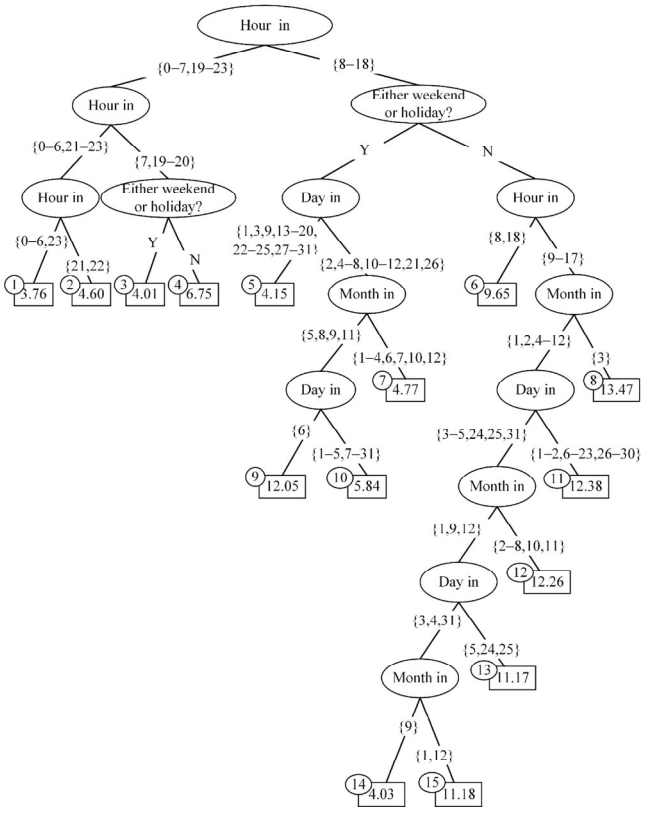

Fig. 9 Identified control strategy and the predicted energy consumption of lighting system (The numbers at leaf nodes are the predicted hourly energy consumption (kWh), e.g., the "13.47" on Node 8 means during 9:00–17:00 on weekdays in March, the lighting energy consumption is 13.47 kWh/h)

Illustrated in Fig. 9, the algorithm selects "hour" as the splitting variable at the root node, and it divides "hour" into two groups, i.e., {0–7,19–23} and {8–18}. If "hour" is in {1–6,21–23}, the electricity usage is relatively low, on the other hand, if "hour" is in {7,19,20}, the electricity usage is in the range between low and high, indicating {7,19,20} is the transition period before or after office hours. The numbers at leaf nodes are the predicted hourly energy consumption (kWh), e.g., the "13.47" on Node 8 means during 9:00–17:00 on weekdays in March, the lighting energy consumption is predicted as 13.47 kWh/h.

It's noted that several abnormal nodes are found in the DT. The predicted energy consumption on Node 9 (Sep 6th, weekend) is abnormally high. And the predicted energy on Node 14 (Sep 3rd and Sep 4th, weekdays) is unusually low. Verified with the building manager, Sep 3rd to Sep 5th in 2015 are holidays and Sep 6th is arranged as a weekday on purpose although it's Sunday.

In order to quantitively validate the accuracy of the identification, the identified operation strategies are used to predict the lighting energy consumption, which would be compared with the measured energy consumption (June, 2016) to validate the accuracy of identified strategy. The validation result is illustrated in Fig. 10. The relative-absolute error (RAE) calculated by predicted data and measured data equals 17.51%, which suggests that the identified operation strategy matches the real operation condition well.

$$\mathrm{RAE} = \frac{\sum_1^n |p_i - a_i|}{\sum_1^n |a_i - \bar{a}|} \tag{6}$$

where n is the number of data points, a_i is the i-th measured value, p_i is the i-th predicted value.

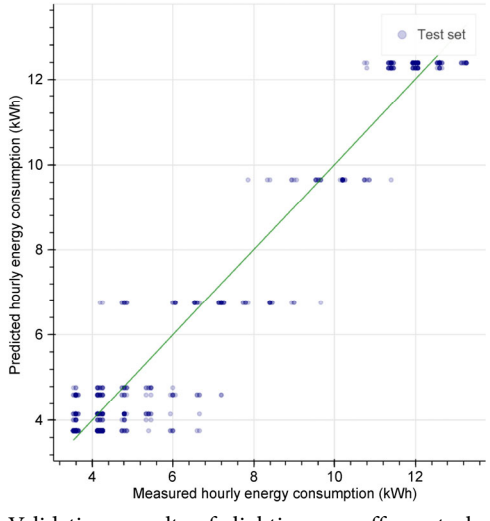

Fig. 10 Validation result of lighting on-off control strategy identification

4.2 Case study 2: chiller sequencing control strategy

The second case study is a commercial building in Shanghai, whose cooling plant consists of 5 chillers (denoted as A, B, C, D and E in the following). Chillers A and B have same rated cooling capacity of 1934 kW, and chillers C, D and E have rated capacity of 3868 kW. Electricity consumption of all five chillers is recorded at 15 min intervals, and the dataset in 2012 is used in this study. To extract on/off status of chillers, a chiller is assumed to be on if its power consumption is larger than 5% of its maximum power usage.

The identification of chillers sequencing control strategy consists of three steps.

First, the maximum cooling capacity of all chillers staged on is used as target variable to construct a decision tree, the result of which is shown in Fig. 11. It's clear that in terms of cooling capacity, the whole cooling season can be roughly divided into five periods: May, June, July to August, September, and October. During each period, the total rated cooling capacity was kept unchanged, although the outdoor temperature changes hour by hour.

Second, the impact of day of month on the choice of total cooling capacity is further explored. As shown in Fig. 12, the transition days during the entire cooling period are: Apr 8th (cooling season starts), May 31st (from period 1 to period 2), Jun 30th (from period 2 to period 3), Aug 31st (from period 3 to period 4), Oct 8th (from period 4 to period 5), and Oct 31st (cooing season ends). The reason period 4 starts from Oct 8th is probably due to the national day holiday (from Oct 1st to Oct 7th), during which the facility managers were off duty. Finally, the possible chiller combinations at each hour during a day are explored, as shown in Fig. 13. During period 5, both chiller A and chiller B are able to provide cooling. However, once a chiller (either A or B) is selected at the start of the day, its status is not changed

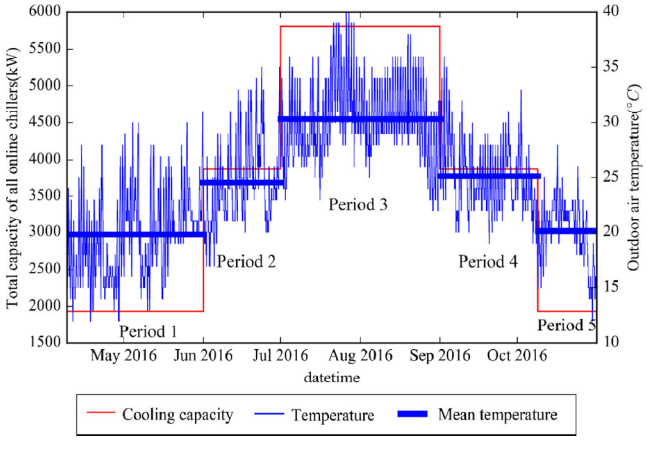

Fig. 11 Outdoor air temperature and total cooling capacity of online chillers

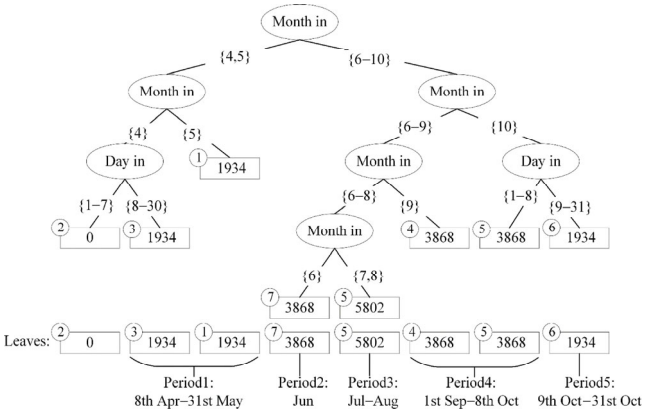

Fig. 12 Dependency of total cooling capacity on time variables

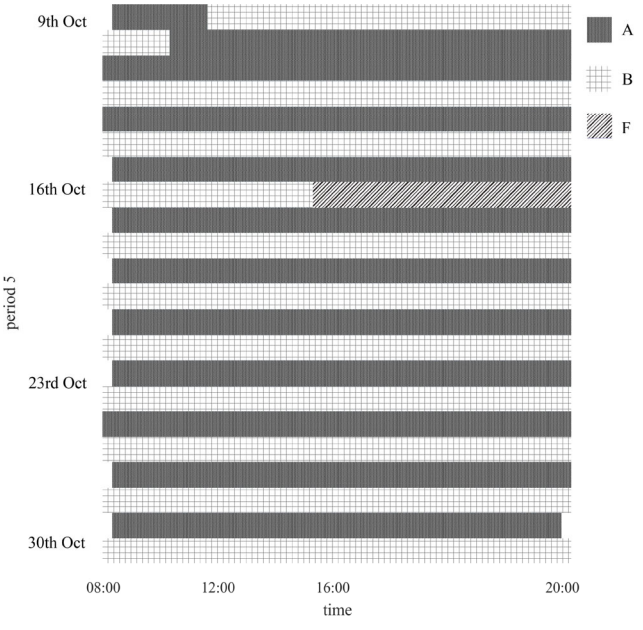

Fig. 13 Sequencing control strategy in period 5

until the end of the day. During the entire Period 5, chillers are switched on/off only in three days (Oct 9th, Oct 10th, and Oct 16th), although ambient temperature fluctuates. Furthermore, the fact that A and B are switched day by day suggests that, the facility manager attempts to distribute the service time to chillers equally.

To sum, the chiller sequencing control strategy for the studied case is identified as the following: first, the total maximum cooling capacity is determined based on month of year; second, different chiller combinations are switched during an operation period to distribute the service time equally; third, during normal working hours in a working day, the on/off status of chillers are kept unchanged. The identification result is validated by an expert in HVAC appliance control who manually identified the chiller operation strategy after on-site survey.

4.3 Case study 3: coordinated control between chiller and pump

The refrigeration system of a hotel building in Shanghai city is studied in Case 3. The system consists of four chillers, six chilled water pumps, and six cooling water pumps. The operational data of these appliances is listed in Table 2. The purpose of this study is to identify the coordinated control strategies between these appliances.

The first step is to discrete the original data. With the number of discretization regions A set as 2 and the number of daily segments W set as 6, the original data is transformed into a set of SAX sentences (shown in Appendix A).

When the SAX transformation is done, WARM approach is applied to the derived data set. During the analysis, the support and confidence are set as 15% and 90% respectively, to keep infrequent but strong rules. As a result, 9 pairs of potentially useful rules are found and they are given in Table 3. Clearly, the antecedent and consequent of the two rules in each pair are exactly reversed, thus shows that the antecedent and consequent nearly occurred at the same time in this year. The first four rules show the association between two components: 2#Chiller and ChilledPumpB1-2. The first pair of rules indicates when the electricity usage of 2#Chiller is low, that of ChilledPumpB1-2 is low as well, and the support of these rules is 0.544. The second pair of rules show that "2#Chiller = high" and "ChilledPumpB1-2 = high" occur at the same time. The total support of these two pairs of rules is 0.988, suggesting 2#Chiller and ChilledPumpB1-2 are turned on/off simultaneously. Similar relationships can be observed for rules 5–16. In Sum, ChilledPumpB1-2, ChilledPumpB1-3 and ChilledPumpB1-4 are dedicated to 2#Chiller, 3#Chiller and 4#Chiller, respectively.

Table 2 Summary of operational data

Number	Name	Acronym	Range (kWh)	Mean (kWh)	Standard variance (kWh)
1	1#Chiller	1C	[0,6.666]	0.001	0.071
2	2#Chiller	2C	[0,1291.666]	186.918	392.040
3	3#Chiller	3C	[0,899.999]	136.847	225.725
4	4#Chiller	4C	[0,300.000]	11.393	43.790
8	'ChilledPumpB1-2'	CPB12	[0,1008.333]	203.593	386.761
9	'ChilledPumpB1-3'	CPB13	[0,1563.636]	739.471	717.225
10	'ChilledPumpB1-4'	CPB14	[0,1458.333]	197.782	484.791
13	'CondPumpB1-1'	CNPB11	[0,966.666]	145.811	337.261
14	'CondPumpB1-2'	CNPB12	[0,400.000]	6.976	45.808
15	'CondPumpB1-3'	CNPB13	[0,26000.000]	707.566	2761.181
16	'CondPumpB1-4'	CNPB14	[0,600.000]	89.103	173.623

Table 3 Rules generated by WARM method

No.	Antecedent	Consequent	Support	Confidence
1	2#Chiller =low	ChilledPumpB1-2 = low	0.544	0.978
2	ChilledPumpB1-2 = low	2#Chiller =low		1
3	2#Chiller=high	ChilledPumpB1-2 = high	0.444	1
4	ChilledPumpB1-2 = high	2#Chiller=high		0.973
5	3#Chiller = low	ChilledPumpB1-3 = low	0.499	0.974
6	ChilledPumpB1-3 = low	3#Chiller = low		1
7	3#Chiller = high	ChilledPumpB1-3 = high	0.488	0.976
8	ChilledPumpB1-3 = high	3#Chiller=high		0.914
9	3#Chiller = low	CondPumpB1-4 =low	0.499	1
10	CondPumpB1-4 =low	3#Chiller = low		0.973
11	3#Chiller = high	CondPumpB1-4 =high	0.453	0.904
12	CondPumpB1-4 =high	3#Chiller = high		0.973
13	4#Chiller = low	ChilledPumpB1-4 = low	0.685	1
14	ChilledPumpB1-4 = low	4#Chiller = low		1
15	4#Chiller = high	ChilledPumpB1-4 = high	0.315	1
16	ChilledPumpB1-4 = high	4#Chiller = high		1
17	CondPumpB1-2=low	CondPumpB1-3=low	0.951	1
18	CondPumpB1-3=low	CondPumpB1-2=low		0.95

The identified coordinated control strategy is verified by the on-site expert and the operator of chilled water system appliances.

4.4 Discussion

The case studies above illustrate how the proposed analysis framework can be applied to identify building operational strategies. It can be found that many of the strategies are schedule based. While lighting system is controlled following a daily schedule, chiller sequencing control is mainly based on day of month and month of year. Although external environment affects the number of chillers required, this influence is not obvious at hourly level. However, although more advanced strategies such as automatic lighting control are not identified in the case studies, the proposed framework is still applicable for buildings operated with those advanced strategies.

5 Conclusion

In this paper, a DM based framework is proposed to identify three types of building operational strategies: on/off control, sequencing control, and coordinated control. This framework mainly consists of two DM methods: Decision Tree (DT) and Weighted Association Rule Mining (WARM). While on/off control and sequencing control strategies are targeted by DT method, coordinated control strategy is identified by WARM method. The framework is validated with three data sets measured by power metering system of buildings in Shanghai. The validation results suggest that the proposed framework is able to automatically and accurately identify lighting on/off control strategy, chiller sequencing control strategy, and coordinated control strategy between chillers and pumps. Implemented on an original software BOSA, this framework is suggested to be used by engineers to automatically identify existing building operation strategies.

Acknowledgements

The authors would like to thank the funding support from Chinese National Science Fund for Young Scholars (No. 51508394).

Appendix A SAX sentence of case study 3 (*D*=24, *W*=6, *A*=2)

2#Chiller	3#Chiller	4#Chiller	CPB12	CPB13	CPB14	CNPB11	CNPB12	CNPB13	CNPB14	Weight
aaaaaa	bbbaab	aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	aaaaaa	aaaaaa	babbbb	0.0067
aaaaaa	bbbbab	aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	aaaaaa	aaaaaa	bbbbbb	0.0470
aaaaaa	bbbbab	aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	aaaaaa	aaaaaa	aabbbb	0.0067
aaaaaa	bbabab	aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	aaaaaa	aaaaaa	bbabbb	0.0067
aaabbb	bbbbbb	aaaaaa	aabbbb	bbbbbb	aaaaaa	aabaaa	aaaaaa	aaaaaa	bbbbbb	0.0067
bbbbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	aaaaaa	aaaaaa	aaaaaa	bbbbbb	0.0336
bbbbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	aaaaaa	aaaaaa	aaaaaa	bbbaab	0.0067
bbbbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	bbbbbb	aaaaaa	aaaaaa	bbbbbb	0.2953
babbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	bbbbbb	aaaaaa	aaaaaa	bbbbbb	0.0134
baabbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	bbbbbb	aaaaaa	aaaaaa	bbbabb	0.0067
babbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	bbbbaa	aaaaaa	aaaaaa	bbbbbb	0.0067
bbbbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	aaaaab	aaaaaa	aaaaaa	bbbbbb	0.0067
baabbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	bbbbbb	aaaaaa	aaaaaa	bbbbbb	0.0067
bbbbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	bbbbbb	aaaaaa	aaaaaa	bbbabb	0.0067
babbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	bbbbbb	aaaaaa	aaaaaa	baaaaa	0.0067
bbbbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	bbbbbb	aaaaaa	aaaaaa	baaabb	0.0067
bbbbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	bbbbbb	aaaaaa	aaaaaa	aaaaab	0.0067
bbbbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	abaaaa	bbbbbb	aaaaaa	aaaaaa	0.0067
bbbbbb	bbbbbb	aaaaaa	bbbbbb	bbbbbb	aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	0.0134
bbbbbb	bbbbaa	aaaaaa	bbbbbb	bbbbaa	aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	0.0067
bbbbbb	aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	0.0201
aaaaaa	aaaaaa	abbbbb	bbaaaa	aaaaaa	abbbbb	aaaaaa	bbaaaa	aaaaaa	aaaaaa	0.0067
aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	bbbbbb	aaaaaa	aaaaaa	aaaaaa	aaaaaa	0.3087
aaaaaa	aaaaaa	baaaaa	aaaaaa	aaaaaa	baaaaa	aaaaaa	aaaaaa	aaaaaa	aaaaaa	0.0067
aaaaaa	aaaaaa	aaaaaa	aaaaaa	aaaaaa	aaaaaa	aaaaaa	aaaaaa	aaaaaa	aaaaaa	0.1544

References

Beghi A, Cecchinato L, Rampazzo M (2011). A multi-phase genetic algorithm for the efficient management of multi-chiller systems. *Energy Conversion and Management*, 52: 1650–1661.

CCIA (2015). China Building Industry Statistical Year Book. China Construction Industry Association. Beijing: China Statistical Press.

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

Dong B, Yan D, Li Z, Jin Y, Feng X, Fontenot H (2018). Modeling occupancy and behavior for better building design and operation—A critical review. *Building Simulation*, 11: 899–921.

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 T (2003). Weighted Association Rule Mining using weighted support and significance framework. Paper presented at ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Fu Y, Li Z, Feng F, Xu P (2016). Data-quality detection and recovery for building energy management and control systems: Case study on submetering. *Science and Technology for the Built Environment*, 22: 798–809.

IEA (1989). World Energy Statistics and Balances: Organisation for Economic Co-operation and Development. International Energy Agency.

Ku K, Jeong S (2018). Building electric energy prediction modeling for BEMS using easily obtainable weather factors with Kriging model and data mining. *Building Simulation*, 11: 739–751.

Lewis R J (2000). An introduction to classification and regression tree (CART) Analysis. Paper presented at the Annual Meeting of the Society for Academic Emergency Medicine.

Li Z, Huang G (2013). Re-evaluation of building cooling load prediction models for use in humid subtropical area. *Energy and Buildings*, 62: 442–449.

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

Magoulès F, Zhao H X, Elizondo D (2013). Development of an RDP neural network for building energy consumption fault detection and diagnosis. *Energy and Buildings*, 62: 133–138.

Miller C, Nagy Z, Schlueter A (2015). Automated daily pattern filtering of measured building performance data. *Automation in Construction*, 49: 1–17.

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

O'Driscoll E, O'Donnell GE (2013). Industrial power and energy metering - a state-of-the-art review. *Journal of Cleaner Production*, 41: 53–64.

Pang Z, Xu P, O'Neill Z, Gu J, Qiu S, Lu X, Li X (2018). Application of mobile positioning occupancy data for building energy simulation: An engineering case study. *Building and Environment*, 141: 1–15.

Sun Y, Huang G, Li Z, Wang S (2013). Multiplexed optimization for complex air conditioning systems. *Building and Environment*, 65: 99–108.

Tibshirani R, Walther G, Hastie T (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, 63: 411–423.

Witten IH, Frank E (1999). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers.

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.

Zhang C, Zhang S (2002). Association Rule Mining. Berlin: Springer.