Mining Big Building Operational Data for Improving Building Energy Efficiency: A Case Study

Cheng Fan[1, 2] and Fu Xiao[*, 2]

[1]Department of Construction Management and Real Estate, Shenzhen University, Shenzhen, China

[2]Department of Building Services Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

[*]E-mail: linda.xiao@polyu.edu.hk; Tel.: (852) 27664194

Abstract

Massive amounts of building operational data are collected and stored in modern buildings, which provide rich information for in-depth investigation and assessment of

actual building operational performance. However, the current utilization of big building operational data is far from being effective due to the gaps between building engineering and advanced big data analytics. Data mining (DM) is a promising technology for extracting previously unknown yet potentially useful insights from big data. This paper aims to explore the potential application of advanced DM techniques for effective utilization of big building operational data. A case study of mining the operational data of an educational building for performance improvement is presented. Decision tree, clustering analysis and association rule mining are adopted to analyze the operational data. The results show that useful knowledge can be extracted for identifying typical building operation patterns, detecting operation deficiencies and spotting energy conservation opportunities.

Keywords: big building operational data; building energy efficiency; decision tree; clustering analysis; association rule mining;

Practical Application

The current utilization of big building operational data in the building industrial is rather limited due to the lack in experience of using advanced big data analytics. This study presents a data mining-based method for analyzing massive building operational data. The case study results validate the efficiency and effectiveness of the method proposed. It can help building professionals to discover valuable insights into building operation patterns and thereby, developing strategies for improving building energy efficiency. The method can be fully realized using the open-source software *R*, which provides great flexibilities in its integration with building automation systems.

1. Introduction

The building sector has become one of the largest energy consumers worldwide. According to the U.S. Energy Information Administration, the Building Sector

(residential and commercial) accounted for 39% of total energy consumption in the United States in 2015 [1]. During the building life cycle, building operation represents the largest portion of electricity use in most developed countries and areas. Buildings shared 74% of the total U.S. retail sales of electricity use in 2015 [2]. In Hong Kong, buildings are responsible for 92% of electricity consumption in 2014 [3]. Building energy efficiency has become a global urgent issue, and attracted great efforts in both academia and industry. To improve the building operational performance, Building Automation Systems (BASs) are usually installed in modern buildings, which facilitate the real-time monitoring, control and energy management. Massive amounts of building operational data are collected and stored in BAS. However, the current utilization of big building operational data is far from being effective due to the lack of suitable methods or tools for data analysis. Conventional methods for analyzing building operational data typically rely on physical principles, statistics and engineering expertise. The inherent

mechanisms of these analytic methods impose great limitations on their abilities in analyzing big data. To address the challenges and opportunities brought by big building operational data, advanced data analytics are urgently needed.

As a promising solution, data mining (DM) technology is renowned for its excellence in knowledge discovery from big data. It has been widely used in various industries, including financial services, retails, health care, and even counter-terrorism [4, 5]. DM techniques can be generally classified into two groups, i.e., supervised and unsupervised DM techniques. Supervised DM performs regression or classification based on the relationships discovered between input and output variables. The knowledge discovered is represented as quantitative or qualitative models. Supervised DM has been applied for energy consumption prediction [6-9] and fault detection and diagnosis [10-13] in the building field. It is noted that supervised DM can barely bring new and exciting knowledge to the building industry, compared with conventional data analytics. The

major reason is that buildings and building systems are well understood. In addition, they usually require the availability of high-quality labeled training data under both normal and abnormal conditions. Such training data can be very hard to obtain in real building operations, e.g., chiller operational data under different faulty conditions. By contrast, unsupervised DM does not require labeled training data and focuses on discovering the intrinsic data structures, correlations and associations. In addition, unsupervised DM requires less domain expertise since there is no need to explicitly pre-define a problem or a mining target, making it more preferable in real applications to discover new knowledge. The knowledge obtained by unsupervised DM is usually represented as data clusters, association rules, and anomalies [4].

This paper presents a case study on extracting useful knowledge from massive building operational data using DM techniques and their potential applications in improving building energy efficiency. The method is developed based on the generic data analytic

framework proposed in our previous study [14]. The main DM techniques adopted are decision trees, clustering analysis and association rule mining. The method has been applied to analyze the data retrieved from an educational building in the Hong Kong Polytechnic University.

2. Research Method

2.1 Research outline

Based on a comprehensive investigation on DM techniques and building operational characteristics, a generic DM-based analytic framework has been developed in our previous work [14]. Four phases are included in this framework, i.e., data exploration, data partitioning, knowledge discovery and post-mining. The data exploration aims to improve data quality and transform data into compatible formats for the implementation of various DM techniques. Data partitioning aims to enhance the reliability and

sensitivity of the knowledge discovered by dividing the whole data sets into several groups according to building operational characteristics. A number of DM techniques are then applied to different data groups separately to extract knowledge at the knowledge discovery phase. Customized post-mining methods are developed to facilitate the process of knowledge interpretation, selection, and applications. The method adopted in this paper is developed from the framework and the main DM techniques used are clustering analysis, decision trees and association rule mining. The following sections present the details.

2.2 Data partitioning

Building operations are highly complex due to the constantly changing indoor requirements and outdoor conditions. It is therefore not wise to treat the building

operational data as a whole for data analysis, as it will downgrade the reliability and sensitivity of knowledge discovered.

Typical building operational data are stored in a two-dimensional data table, in which each column represents a variable and each row stores measurements sampled at the same time step. Data partitioning refers to the process of dividing the entire data table into several subsets, each containing a number of rows.

It is found that two types of methods are suitable for partitioning building operational data. The first is to treat each row as an observation and then grouping observations based on their similarities. Clustering analysis is one of the most suitable methods to perform this task. The aim of clustering analysis is to divide data into several clusters while maximizing the within-cluster similarities and minimizing the between-cluster similarities. There are three general types of clustering methods, i.e., hierarchical, partitioning and density-based methods [4]. The similarities among observations can be

evaluated using various distance measures, such as the Euclidean distance and Cosine distance. It is worth mentioning that due to the curse of dimensionality, the distance-based similarity measures may become meaningless when the variable number is large [15]. Therefore, users may have to select a small subset of variables as inputs for clustering analysis. This subset should be able to reflect the changes in building operations. The main drawback of this method is that the result lacks interpretability, as the only output is the clustering membership. Further analysis has to be carried out if users want to know the data characteristics in each cluster.

The other method is to partition the data according to one representative variable, which can describe the operation characteristics and is also a major concern of building professionals, e.g., the building power consumption and the building cooling load. The decision tree method, which is one of the most widely used predictive data mining techniques, can be applied for this task. A decision tree model adopts a tree structure to

present the logic flow. The aim is to predict the value of a target variable based on inputs. A decision tree model has three main components, i.e., the root, internal and terminal nodes. An input variable is selected at each root and internal nodes to enhance the prediction performance. Commonly used metrics for performance evaluation include the Gini impurity, information gain, misclassification rate and variance [4]. The prediction outcomes are shown in terminal nodes and the whole model can be easily interpreted as decision rules. An example of decision tree model is shown in Fig. 1. The model describes fictional relationships between gender, hair and shoes. Node 1 is the root node and the splitting variable selected is the shoe type. Node 3 is called internal node. Nodes 2, 4 and 5 are called terminal nodes, which present the prediction result of gender. Node 2 indicates that the gender should be 100% *Female* if one wears high heels. Node 5 states that the gender is 100% *Male* if one wears sneaker and has short hair. Node 4 indicates that if one wears sneakers and has long hair, the gender can be

either *Female* or *Male* with possibilities of 60% and 40% respectively. The decision tree model is highly interpretable and provides detailed clues on how to partition the data. The decision method is applied for the data partitioning task in this study.

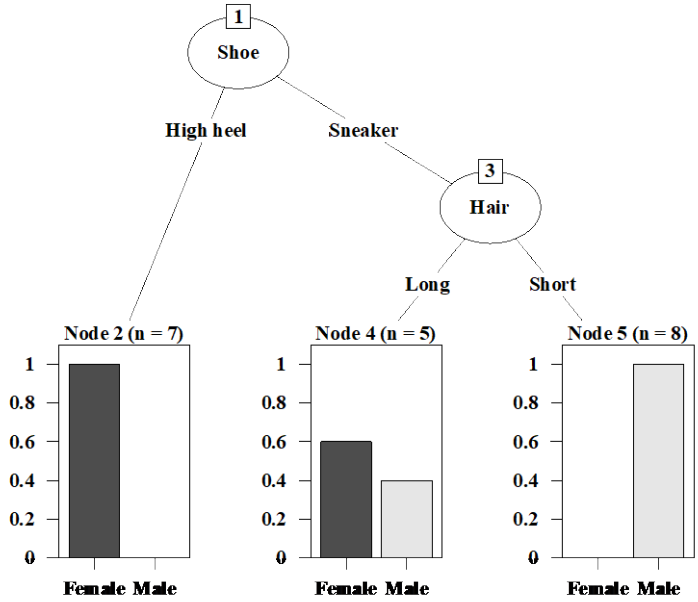

Fig. 1. An example decision tree model

2.3 Knowledge discovery

Investigating the relationships between different variables is the main approach for knowledge discovery. Association rule mining is a popular method to mine associations

among variables. An association rule A → B states that if *A* happens, then *B* happens, where *A* is the antecedent and *B* is the consequence. Two thresholds are typically used for rule generation, i.e., support and confidence. The support is the joint probability of *A* and *B* both happening while the confidence is the conditional probability of *B* given *A*. The number of association rules obtained decreases with the increase in the support and confidence thresholds. Users may use some statistics for selecting potentially interesting rules. For instance, the lift value defines the ratio between the rule confidence and the support of the consequence [4]. A lift larger than 1 indicates that the presence of consequence is positively affected by the presence of antecedence and vice versa. A lift of 1 indicates that the antecedent and consequent are independent from each other.

Conventional association rule mining algorithms only work with categorical variables. The majority of building operational data is numeric and therefore, discretization becomes necessary. Data discretization for building operational data is a challenging

problem, as variables usually have their own behavior and the optimal discretization methods are hard to develop without in-depth domain knowledge. Improper discretization usually leads to information loss and may severely downgrade the mining performance. Quantitative association rule mining algorithms are therefore proposed so that both numeric and categorical variables can be mined directly without manual discretization.

This study adopts the QuantMiner [16] as the mining algorithm. If the variable is numeric, an interval is automatically identified considering the rule gain and the coverage of the interval identified. The interval identified is then used to create categorical values for rule generation. The rule gain is calculated using Eq. 1, where *MinConf* refers to the minimal confidence threshold. Genetic algorithm is applied to identify the interval by maximizing a fitness function, as shown in Eq. 2, where A_{num} is the number of numeric variables in the rule; I_{Ai} is the interval identified of A_i; $size(A_i)$ is

the range of A_i; $size(I_{Ai})$ is the length of the identified interval. The algorithm prefers to select rules with large gains and small intervals.

$$Gain(A \rightarrow B) = Support(A, B) - MinConf \times Support(A) \qquad \text{(Eq. 1)}$$

$$Fitness(A \rightarrow B) = Gain(A \rightarrow B) \times \prod_{A_i \in A_{num}} \left[1 - \frac{size(I_{A_i})}{size(A_i)}\right]^2 \qquad \text{(Eq. 2)}$$

3. Case study

3.1 Description of building, system and data

An educational building in the Hong Kong Polytechnic University is selected for analysis. It mainly serves as offices and classrooms. The gross floor area is approximately 11,000m^2, of which about 8,500m^2 are air-conditioned spaces.

The building operational data under concern recorded the operating conditions of the Heating, Ventilation and Air-Conditioning (HVAC) waterside system at the interval of 30-minute. The chiller plant contains 4 water-cooled chillers (denoted as CH-1 to CH-4) and 4 cooling towers (denoted as CT-1 to CT-4). Chillers are connected in parallel and

the chilled water is distributed using 6 primary chilled water pumps (denoted as PCHWP-1 to 6) and 6 secondary chilled water pumps (denoted as SCHWP-1 to 6). The condenser water is circulated between chillers and 4 cooling towers using 6 variable-speed pumps (denoted as CDWP-1 to 6). One-year data retrieved from the BAS (from January 2015 to December 2015) are analyzed. The data have 17,110 observations of 113 variables, including almost all the major variables of the HVAC waterside system, e.g., temperature, flow rate, pressure and on/off signals.

3.2 Data partitioning using decision tree method

Building cooling load, which is sensitive to the outdoor and indoor conditions, is an essential variable in building energy management. It can be used as an indicator of different building operation patterns. As introduced in section 2.2, the decision tree method is adopted for data partitioning. In this study, the building cooling load is

considered as the output variable and the time variables, such as the *Year*, *Month*, *Day*, *Hour*, *Minute* and *Day type*, are used as the input variables. The indoor variables, such as the occupant number, are not used as inputs because, firstly, those data are not available due to the lack of measurement instruments; secondly, they are not necessary considering that the time and day type can also describe how people use the spaces for educational buildings.

The decision tree model constructed is shown in Fig. 2. The model selects the *Month*, *Hour* and *Day type* as splitting variables. Starting from Node 1, the model first picks the *Hour* as the splitting variable and the splitting criterion is {0, 1, 2, 3, 4, 5, 6, 7, 23} and {8 to 22}. The result matches our domain knowledge as it corresponds to the non-peak and peak hours. The lectures normally start at 8:30am and end at 9:30pm. Node 3 divides the data based on the *Day type* and the partitioning is made based on {Monday to Saturday} and {Sunday}. It should be noted at many classes for part-time students

and academic events are scheduled in this building on Saturdays and therefore, the cooling load on Saturdays is very similar to that on a typical weekday. Node 5 selects *Month* as the splitting variable and the two splitting sets are {1, 2, 3, 4, 12} and {5, 6, 7, 8, 9, 10, 11}. The first set corresponds to the cooler and less humid seasons while the second refers to the hotter and more humid seasons in Hong Kong.

The decision tree model developed provides evident clues on data partitioning. Rather than dividing the whole data into 4 data groups according to the terminal nodes (i.e., Nodes 2, 4, 6 and 7), the splitting criteria generated at Nodes 1, 3 and 5 are used together to partition the data in a more comprehensive manner. As a result, the entire data sets are partitioned into 8 groups, as shown in Table 1. Fig. 3 presents the distribution of building cooling load in each data partition. It is apparent that the building cooling load in each group presents different distributions, especially when it belongs to peak hours. The cooling loads during non-peak hours during Mondays to

Saturdays and Sundays in the same cool or hot season are generally the same, e.g. comparing Group 1 and Group 3, Group 5 and Group 7 in Fig. 3. It is worth mentioning that data distributions in Groups 4, 5 and 7 are quite close, which states that the cooling load during peak hours on Sundays in cool seasons is similar to that during non-peak hours in hot seasons. Other data partitioning approaches (e.g., clustering analysis) may group those observations into one group. The decision tree method therefore provides more detailed partitioning results which would improve the sensitivity and reliability of the knowledge discovered.

Table 1. Details on eight data groups

Groups	Month	Day type	Hour
1	{1,2,3,4,12}	{Monday to Saturday}	{0,1,2,3,4,5,6,7,23}
2	{1,2,3,4,12}	{Monday to Saturday}	{8 to 22}
3	{1,2,3,4,12}	{Sunday}	{0,1,2,3,4,5,6,7,23}

4	{1,2,3,4,12}	{Sunday}	{8 to 22}
5	{5 to 11}	{Monday to Saturday}	{0,1,2,3,4,5,6,7,23}
6	{5 to 11}	{Monday to Saturday}	{8 to 22}
7	{5 to 11}	{Sunday}	{0,1,2,3,4,5,6,7,23}
8	{5 to 11}	{Sunday}	{8 to 22}

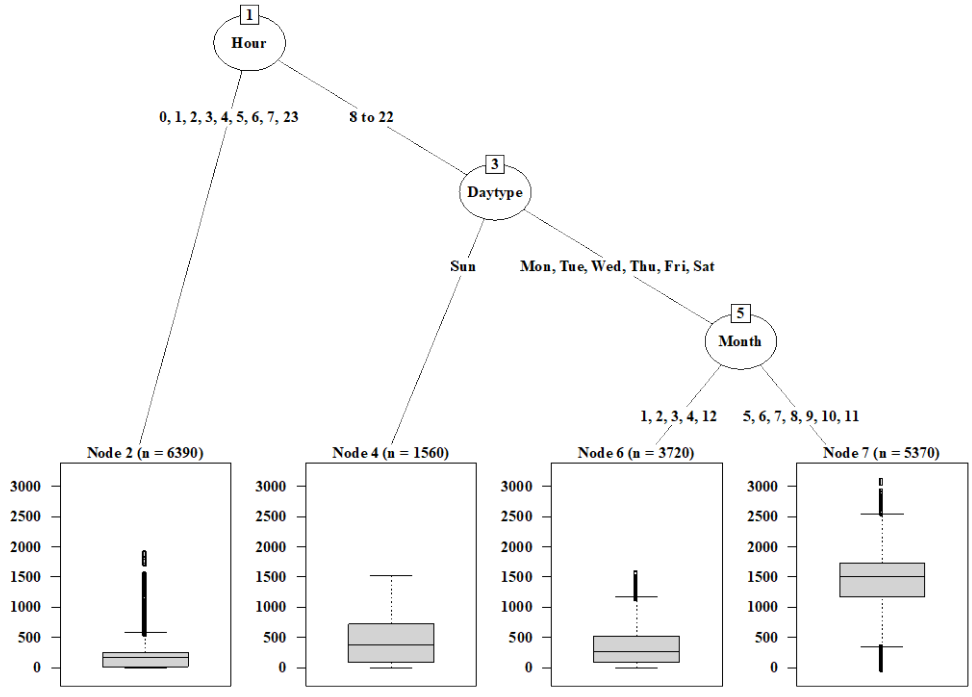

Fig. 2. Decision tree model for building cooling load

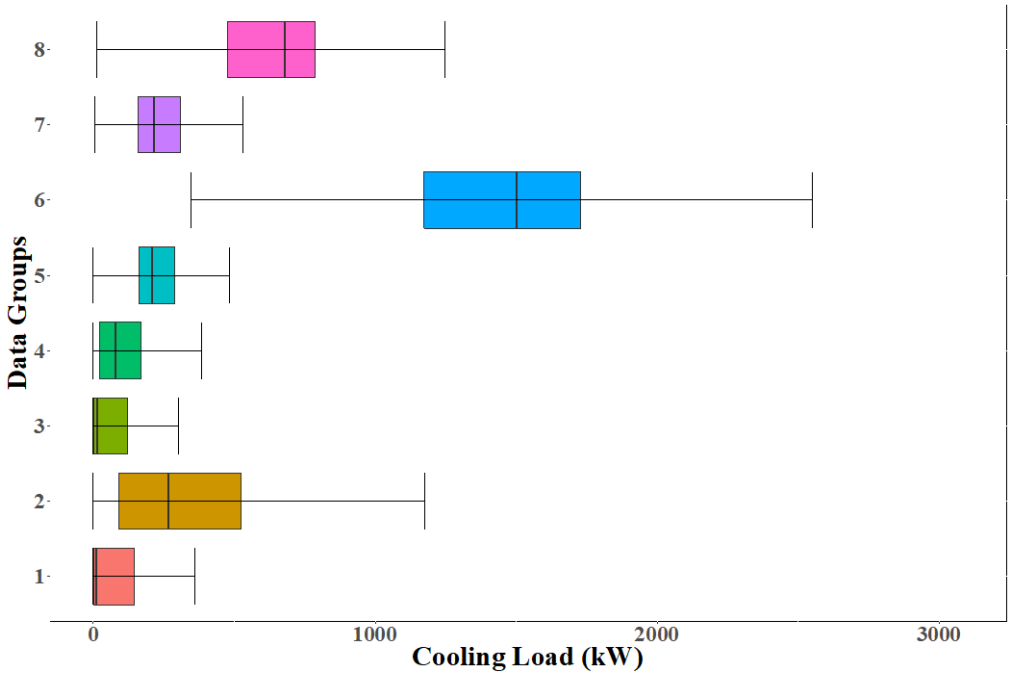

Fig. 3. Boxplots of building cooling load in each data group

3.3 Knowledge discovery using quantitative association rule mining

As introduced in section 2.3, the QuantMiner algorithm is adopted to discover the associations in each data group separately. For the convenience of rule interpretation, both sides of the rule, i.e. the antecedent and the consequence, are constrained to have one variable only. The parameters for the genetic algorithm are set as follows: 250 as

population size, 100 as iteration number, 50% as crossover rate and 40% as mutation rate. These parameters are set according to the suggestions in [16]. The support and confidence thresholds are set as 5% and 90% respectively. In general, the confidence threshold should be set no less than 80% to ensure the quality of association rules. The support values can be set according to the user's actual need. A small support threshold leads to the discovery of less frequent associations. It can be used to discover atypical associations in building operations. However, a smaller support threshold will cause a dramatic increase in the number of association rules obtained, which makes the post-mining phase more time-consuming. Building professionals should also pay attention to the inherent support of a class when trying to derive association rules containing that class. If the inherent support of a class is small, the support threshold should be set even smaller so that association rules containing that class can be discovered.

Taking Mondays to Saturdays in hot seasons (i.e., Group 5 and 6) as examples, 199 and 161 quantitative association rules are obtained respectively. The majority of the rules obtained are in accordance with domain expertise and Table 2 presents 3 example rules. The first rule states that if the number of running chillers is 0, then the total condenser water flow will range from 0.0 to 1.0 *l/s*. The rule confidence is quite high but not 100%. This is because the water flow sensor may have recorded some values slightly smaller than 0 or larger than 1 due to the measurement precision problem or the data transmission problem. The latter two rules specifying the idle condition of CT-2 and CH-2 also agree with domain expertise. It should be mentioned that such rules could be used as a knowledge database, which can be further applied to detect anomalies in new observations. Meanwhile, some rules are not in accordance with expectation. These rules can be directly applied for detecting anomalies, faults and deficiencies in building

operations and thereby, identifying energy conservation opportunities. The details are discussed in the following section.

Table 2. Example quantitative association rules in data groups 5 and 6

No.	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Data group
1	CH_No = 0	CDW_Flow in [0.0, 1.0]	87.9	98.8	1.2	5
2	CT2_Status = Off	CT2_MotorSpeed in [0.0, 0.6]	89.6	100	1.1	5
3	CH2_status = Off	CH2_CHW_Flow in [-0.1, 0.3]	73.4	99.5	1.4	6

4. Applications

4.1 Identification of energy conservation opportunities

Chilled water and condensing water distribution system

Two examples rules presented in Table 3 indicate that when one chiller is switched on,

its chilled water and condensing water flow rates become nearly constant. By checking the actual motor speed of PCHWP and CDWP, it is found out that the motor frequency was maintained at 40Hz during operation, indicating that the energy saving potential of variable speed operation was not realized. The insights obtained helps to spot the energy conservations in actual operations, as control strategy should be developed to optimize the pressure set-point for pump speed control according to the actual cooling load and weather conditions.

Table 3. Example quantitative associations in chilled water and condensing water flow rates

No.	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Data group
1	CH1_Status = On	CH1_CHW_Flow in [47.1, 51.3]	60.1	99.5	1.7	6

2	CH2_Status = On	CH2_CDW_Flow in [46.8, 51.9]	26.1	99.7	3.8	6

Chiller control strategy

The rules in Table 4 describe the supplied chilled water temperature when one chiller is switched on. The intervals identified for the chilled water supply temperature are quite narrow. It turns out that the supplied chilled water temperature set-point was set fixed as 7^{o}C. Considering that the chilled water supply temperature has a huge impact on the chiller power consumption [17], it is suggested to develop a temperature reset scheme to regulate the chilled water supply temperature.

Table 4. Example quantitative associations in chiller operation

No.	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Data group

1	CH1_Status = On	CH1_CHW_ST in [6.8, 7.3]	60.0	99.4	1.7	6
2	CH2_Status = On	CH2_CHW_ST in [6.8, 7.8]	26.0	99.2	3.8	6

Cooling tower control strategy

The rules in Table 5 indicate that the cooling tower fan speed was maintained at around 35Hz during operations, which indicates that the energy saving potential through variable speed control was not achieved. An optimal condenser inlet water temperature set-point reset scheme should be developed to provide fan speed set-points according to the ambient and working conditions with the aim of minimizing the overall energy use of chillers and cooling tower fans.

Table 5. Example quantitative associations in cooling tower operation

No.	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Data group
1	CT1_Status = On	CT1_MotorFrequency in [35.0, 35.6]	84.7	99.5	1.2	6
2	CT2_Status = On	CT2_MotorFrequency in [35.2, 35.7]	82.5	99.6	3.2	6

4.2 Evaluation on HVAC operational performance

The HVAC operational performance can be evaluated using the system coefficient of performance (COP), which equals to the ratio between the building cooling load and the power consumption of the chiller plant (i.e., including chillers, chilled water and condensing water pumps, cooling towers). The equal-width binning method is applied to discretize the system COP into three classes for performance evaluation, namely as

Poor, *Medium* and *Good*. If none of the chillers are switched on, the system COP is represented as *Idle*. Fig. 4 presents the distribution of system COP. The red vertical dashed lines represent the cutting points for data discretization. The cutting points are selected as 2.3 and 3.5. Example quantitative associations for performance evaluation are shown in Table 6 and details are presented as follows.

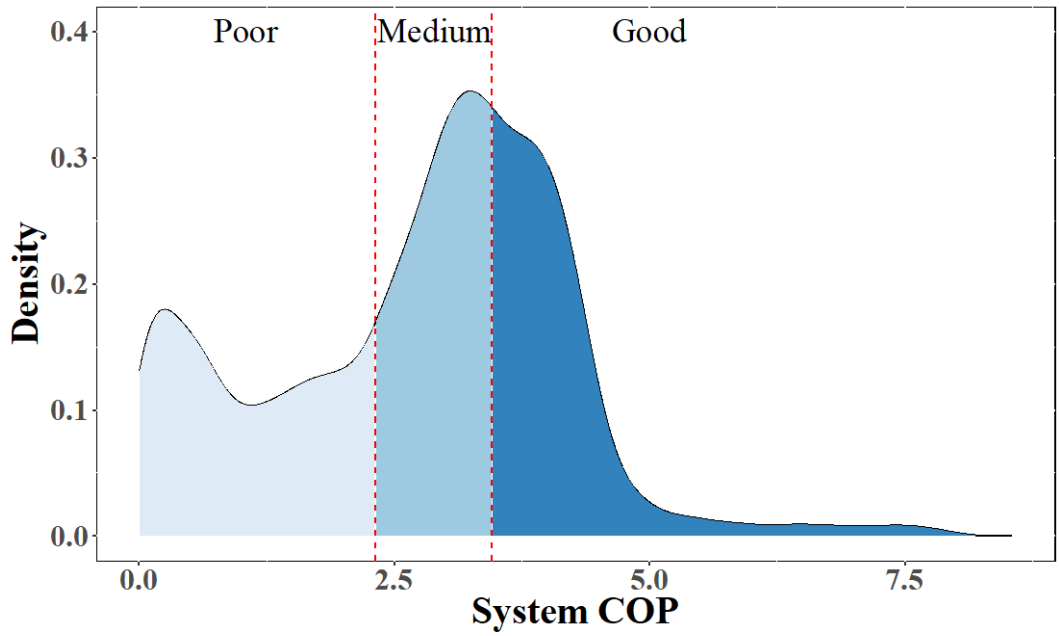

Fig. 4. Density plot of system COP

Table 6. Quantitative associations for system performance evaluation

No.	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Data group
1	PLR in [0.1, 0.2]	Performance = Poor	21.8	89.5	2.2	5
2	PLR in [0.1, 0.2]	Performance = Poor	12.1	89.1	3.0	6
3	PLR in [0.9, 1.0]	Performance = Good	5.6	98.4	5.3	5
4	PLR in [0.5, 0.7]	Performance = Good	31.2	94.8	1.5	6
5	Cooling Load in [459.0, 632.5] kW	Performance = Good	6.0	88.7	4.7	5

6	Cooling Load in [951.4, 1366.2]	Performance = Good	23.1	90.7	1.4	6
7	CH_1 = Off	Performance = Good	52.5	81.1	1.3	6

Identification of major influential factors to system performance

The first two rules in Table 6 describe that when the PLR is between 0.1 and 0.2, the system performance is *Poor*. This is in accordance with domain expertise as low PLRs usually lead to poor energy efficiency in chiller operations. The third and fourth rules describe the PLR intervals identified in Groups 5 and 6 when the system performance is *Good*. As expected, these two PLRs are much higher than the intervals identified in the first two rules. It is noted that when the system performance is *Good*, the PLRs identified in Group 6 (i.e., between 0.5 and 0.7) is lower than those in Group 5 (i.e.,

between 0.9 and 1.0). It turns out that a small air-cooled chiller is used to fulfill the cooling load demand during non-peak hours (i.e., data in Group 5), while the other three water-cooled chillers are used during peak-hours (i.e., data in Group 6). These two rules are actually justifications for the claim that water-cooled chillers present higher energy efficiency than air-cooled chillers. Rules No. 5 and 6 depict the quantitative associations between cooling load and system performance. They shed insights into the cooling load intervals which can be fulfilled by the existing system with better energy efficiency.

Identification of less energy-efficient components

Rule No. 7 in Table 6 presents an interesting association between system components and system performance. It states that if CH-1 is switched off, then the system performance is *Good*. This rule initiates a hypothesis that CH-1 is less energy-efficient

compared to the other 2 water-cooled chillers (note that CH-4 is not in operation in Group 6).

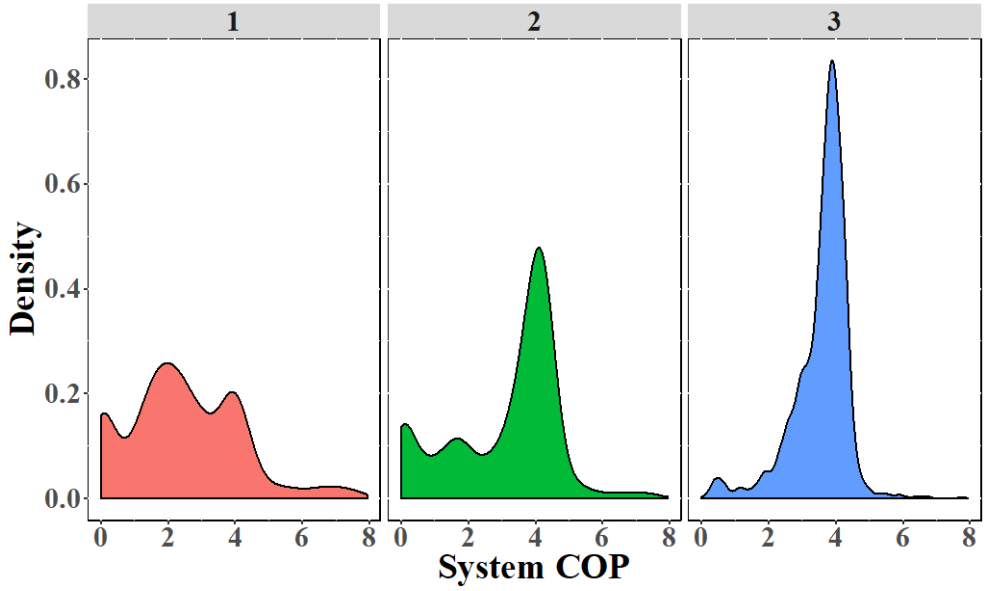

Fig. 5. Density plot of system COP when CH-1 to 3 is in operation

To further investigate on this hypothesis, Fig. 5 is drawn to compare the system COP when CH-1, CH-2 and CH-3 are in operation alone. It is evident that using CH-2 or CH-3 results in better system performance than using CH-1. As a more concrete proof, the two-sample T-test is performed. Given a confidence level of 95%, the P-values obtained show that the null hypothesis (i.e., there is no difference between two-sample

means) fails to be rejected considering CH-2 and CH-3, while getting rejected between CH-1 and CH-2, CH-1 and CH-3. This further confirms that the operation of CH-1 leads to less energy-efficient operations.

5. Conclusion

Massive amounts of building operational data are being collected and stored in modern buildings. How to effectively and efficiently transform big building data into useful insights and actionable measures for better building energy management is an urgent challenge to tackle.

In this paper, we present the potential application of data mining (DM) in discovering useful knowledge from massive building operational data through a case study. The method is developed from the generic DM-based analytic framework proposed in our previous work. The decision tree method is applied to develop models on building

cooling load and thereby, providing comprehensive guidelines on data partitioning. Once the data are partitioned into different groups, association rule mining is applied separately to derive associations. The quantitative association rule mining is selected as the main tool since building operational data contains both continuous and categorical variables. Association rule mining is an unsupervised DM technique, which requires no training data and little prior knowledge on building operations. It therefore provides the greatest flexibilities in real practice and has the ability to discover previously unknown knowledge.

The research results show that the method proposed can extract valuable knowledge from building operational data with high efficiency and effectiveness. Typical operation patterns and control strategies of HVAC systems have been discovered while revealing opportunities for enhancing building energy efficiency. It should be mentioned that the data used in this study might not be large enough to be quoted as ‘big data’. The

emphasis of this study is the analytic method, which can be scalable to more complex datasets. Further studies will be performed to exploit practical applications in building energy management and validate the mining performance when larger datasets are available.

6. Acknowledgements

The authors gratefully acknowledge the support of this research by the Research Grant Council (RGC) of the Hong Kong SAR (152181/14E), the Hong Kong Polytechnic University and Shenzhen University.

References

[1] U.S. Energy Information Administration. *Monthly energy review*. January 2017. http://www.eia.gov/totalenergy/data/monthly/index.php

[2] U.S. Energy Inforamtion Administration. *Annual energy outlook 2016*. September 2016. https://www.eia.gov/outlooks/aeo/tables_ref.cfm

[3] Electrical and Mechanical Services Department of the Hong Kong SAR Government. *Hong Kong Energy End-Use Data*. 2016.

[4] Han JW, Kamber M, Pei J. *Data mining: Concepts and techniques*. 3rd ed. Massachusetts: Morgan Kaufmann, 2012.

[5] Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications-A decade review from 2000 to 2011. Expert Syst Appl 2012; 39: 11303-11311.

[6] Fan C, Xiao F. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energ 2014; 127: 1-10.

[7] Zhao DY, Zhong M, Zhang X, Su X. Energy consumption predicting model of VRV (variable refrigerant volume) system in office buildings based on data mining. Energy 2016; 102: 660-668.

[8] Le Cam M, Daoud A, Zmeureanu R. Forecasting electric demand of supply fan using data mining techniques. Energy 2016; 101: 541-557.

[9] Zeng YH, Zhang ZJ, Kusiak A. Predictive modeling and optimization of a multi-zone HVAC system with data mining and firefly algorithms. Energy 2015; 86: 393-402.

[10] Capozzoli A, Lauro F, Khan I. Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Syst Appl 2015; 42: 4324-4338.

[11] Fan C, Xiao F, Madsen H, Wang D. Temporal knowledge discovery in big BAS data for building energy management. Energ Buildings 2015; 109: 75-89.

[12] Xiao F, Fan C. Data mining in building automation system for improving building operational performance. Energ Buildings 2014; 75: 109-118.

[13] Du ZM, Fan B, Jin XQ, Chi JL. Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. Build Environ 2014; 73: 1-11.

[14] Fan C, Xiao F, Yan CC. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Automat Constr 2015; 50: 81-90.

[15] Kriegel HP, Kroger P, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans Knowl Discov Data 2009; 3: Article 1.

[16] Salleb-Aouissi A, Vrain C, Nortet C. QuantMiner: A genetic algorithm for mining quantitative association rules. IJCAI 2007；1035-1040.

[17] Yan CC, Wang SW, Xiao F, Gao DC. A multi-level energy performance diagnosis method for energy information poor buildings. Energy 2015; 83: 189-203.

Fig. 1. An example decision tree model

Fig. 2. Decision tree model for building cooling load

Fig. 3. Boxplots of building cooling load in each data group

Fig. 4. Density plot of system COP

Fig. 5. Density plot of system COP when CH-1 to 3 is in operation