https://doi.org/10.1016/j.autcon.2014.12.006

A Framework for Knowledge Discovery in Massive Building Automation Data and Its Application in Building Diagnostics

Cheng Fan, Fu Xiao* and Chengchu Yan

Department of Building Services Engineering, The Hong Kong Polytechnic University

Hung Hom, Kowloon, Hong Kong

*Corresponding Author: Tel.: + 852 27664194; Fax: +852 2765 7198;

Email address: linda.xiao@polyu.edu.hk

Abstract

Building Automation System (BAS) plays an important role in building operation nowadays. A huge amount of building operational data is stored in BAS; however, the data can seldom be effectively utilized due to the lack of powerful tools for analyzing the large data. Data mining (DM) is a promising technology for discovering knowledge hidden in large data. This paper presents a generic framework for knowledge discovery in massive BAS data using DM techniques. The framework is specifically designed considering the low quality and complexity of BAS data, the diversity of advanced DM techniques, as well as the integration of knowledge discovered by DM techniques and domain knowledge in the building field. The framework mainly consists of four phases, i.e., data exploration, data partitioning, knowledge discovery, and post-mining. The framework is applied to analyze the BAS data of the tallest building in Hong Kong. The analysis of variance (ANOVA)

method is adopted to identify the most significant time variables to the aggregated power consumption. Then the clustering analysis is used to identify the typical operation patterns in terms of power consumption. Eight operation patterns have been identified and therefore the entire BAS data are partitioned into eight subsets. The quantitative association rule mining (QARM) method is adopted for knowledge discovery in each subset considering most of BAS data are numeric type. To enhance the efficiency of the post-mining phase, two indices are proposed for fast and conveniently identifying and utilizing potentially interesting rules discovered by QARM. The knowledge discovered is successfully used for understanding the building operating behaviors, identifying non-typical operating conditions and detecting faulty conditions.

Keywords: building automation system, data mining, building energy performance, building diagnostics

Nomenclature

AD	Abnormality Degree
AHU	Air-handling Units
BAS	Building Automation System
CDWP	Condenser Water Pumps
CT	Cooling Towers
ELTG	Essential Power and Lighting
MV	Mechanical Ventilation
NLTG	Normal Power and Lighting

PAU	Primary Air-handling Units
PCHWP	Primary Chiller Water Pumps
PD	Plumbing and Drainage
SCHWP	Secondary Chilled Water Pumps
SD-Lift	Standard Deviation of Lift
Temp_rtn_ch	Returned Chilled Water Temperature
Temp_sup_ch	Supplied Chilled Water Temperature
VTS	Vertical Transportation System
WCC	Water-cooled Chillers

Introduction

Modern buildings, particularly the public and commercial buildings, are equipped with Building Automation Systems (BASs) for real-time monitoring and controlling of the complicated services systems, including air conditioning, lighting, vertical transportation system, security systems and etc. BASs are the products of modern information technology, computing science and control theory. They are essentially networks of a range of hardware devices (e.g., servers, workstations, digital controllers and sensors) and software (e.g., building energy management programs and network communication protocols). A recent report showed that the potential energy savings from the adoption of advanced building automation technology might reach 22% by 2028 for the European building sector [1]. The savings are amazing considering that the building sector is responsible for approximately 32% of total final energy consumption and 40% of primary energy consumption in

most countries [2]. The functionalities of BASs determine the building operational performance to a large extent. To fulfill the functions of BAS, real-time operational data are collected and stored at short intervals (from tens of seconds to several minutes) which results that a tremendous amount of building operational data is available in BASs. The amount of the BAS data keeps increasing along the building life cycle. However, the big sets of data in BASs are not fully utilized due to the lack of advanced data analysis techniques and tools. Today's BASs can only perform rather simple data analysis, such as historical data tracking, moving averages and benchmarking. In the last decade, more sophisticated tools were developed and installed in BASs owing to the fruitful research and development efforts made on advanced optimization and diagnostics of buildings [3, 4]. However, those tools only take advantage of a small amount of data in BASs, and focus on the problems associated with a component or subsystem. Meanwhile, BAS data usually contains a substantial number of missing values and outliers. If those data are used in data analysis, they would ruin the analysis process and the results obtained would hardly be reliable. The building automation industry needs advanced techniques and powerful tools to analyze the massive operational data in BASs so as to understand, evaluate and improve the building operational performance.

Data mining (DM) is a promising technology, which provides new approaches to handling massive and complex data. MIT Review considered DM as one of the top 10 emerging technologies that will change the world [5]. DM has been successfully applied in various fields, such as retails, telecommunication, and financial services [6]. DM techniques can be roughly classified into two categories, i.e., supervised learning and unsupervised learning techniques. Supervised learning aims to establish the relationship between the outputs and inputs by learning from the historical data. By

contrast, unsupervised learning is not guided by an explicit mining target, and its aim is to identify the underlying and unknown data structures or associations between variables. Interests in the use of DM in the building field are increasing in recent years. DM techniques have found their strengths in three areas of the building field, i.e., prediction [7-9], fault detection and diagnosis [10-12], and control optimization [13-15]. However, the potential of DM in the knowledge discovery in massive BAS data has not been fully exploited. Previous research relied heavily on domain knowledge and mainly used supervised learning techniques. The problems were usually predefined and only a small subset of BAS data was used. For instance, in the development of the prediction model for the chiller power consumption [16], inputs to the model, e.g., the supply and return temperatures of chilled water, and the supply and return temperatures of condenser water, were selected in advance, since domain expertise tells us that these variables are the most influential variables to chiller power consumption. Even though the developed models may have higher accuracy owing to the use of domain expertise and advanced DM techniques, knowledge being discovered underlying the massive BAS data is limited.

On the one hand, although DM technology brings valuable opportunity to effective utilization of massive BAS data, the application of DM techniques in the building field faces great challenges. DM itself cannot tell the value or the significance of the knowledge discovered, and domain knowledge is still needed to interpret the knowledge for practical applications. The knowledge discovered by DM is usually enormous and may be in various forms, such as clusters, association rules, statistics, and predictive models. Meanwhile, advanced DM techniques are constantly emerging. It is not easy for building professionals to catch up the progress of DM technology. How to select the most suitable

DM techniques and how to select practically valuable knowledge are two big challenges. It is not wise to attempt individual DM technique and interpret knowledge discovered on a case-by-case basis. To enable the entire building automation industry to benefit from the advanced DM technology, a generic framework for knowledge discovery in massive BAS data using DM techniques is needed. The framework should also take the poor quality of BAS data, which always contains a large number of missing values and outliers, into consideration. This paper presents a generic framework for knowledge discovery in massive BAS data using DM techniques. It is specifically designed to address all the challenges above-mentioned. The framework mainly consists of four phases, i.e., data exploration, data partitioning, knowledge discovery, and post-mining. It is expected that software tools compatible with modern BASs can be developed based on the framework. The framework is applied to analyze the BAS data of the tallest building in Hong Kong. Its values in facilitating building diagnostics are impressive.

2. Description of the framework

The framework developed is shown in Fig. 1, which mainly includes four phases. Data exploration consists of two tasks, i.e., data preprocessing and visualization. Data preprocessing aims to enhance the data quality and transform the data to suitable formats as required by DM techniques. Visualization helps the users to visually gain preliminary understanding about the data. Data partitioning aims to identify the typical building operating patterns so that the large BAS datasets can be partitioned into several subsets. It is important for enhancing the efficiency and reliability of the knowledge discovery by separately mining the data in each pattern. Knowledge discovery may adopt a number of DM techniques, such as association rule mining, clustering analysis, sequential pattern

mining, ensemble learning, classification and regression, to discover the hidden knowledge. Post-mining aims to select, interpret, and make use of knowledge discovered. This study develops a novel method for selecting potentially useful association rules from a large number of rules discovered, which can significantly reduce the time needed to interpret rules using domain knowledge. As last, the selected knowledge can be used for specific tasks including performance evaluation, abnormality detection and control optimization. In this paper, the application in building diagnostics is investigated. The following sections explain the details of each phase and the suitable DM techniques and algorithms.

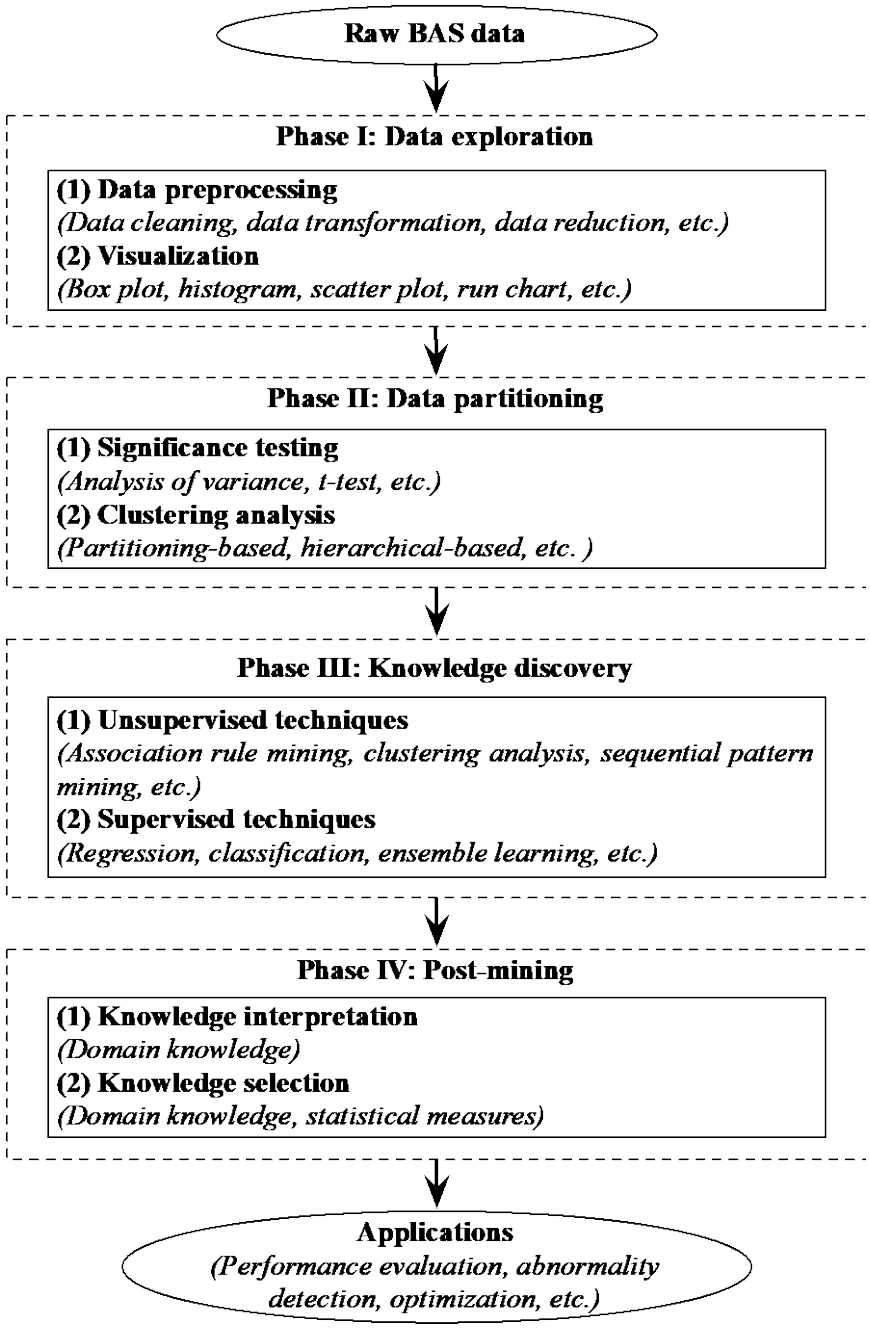

Fig.1 Framework for mining BAS data using DM techniques

2.1. Data exploration

The two main tasks in the data exploration phase are data preprocessing and data visualization. Data preprocessing is an essential step in a knowledge discovery process and it may take 80% of the total DM efforts [17]. Data preprocessing involves data cleaning, data transformation and data reduction. Data cleaning aims to enhance data quality considering missing values, inconsistencies and outliers,

which widely exist in BAS data. Missing value may be caused by the sensor malfunctions or signal transmission errors. Moving average, imputation, and inference-based methods can be used to fill in the missing values [18]. Inconsistencies refer to the differences in the data scales or units, and unmatched records in different data sources. It can be solved using data fusion schemes [19] or physical redundancy. Outliers are those records that deviate from their true values, which can be detected using statistical methods [20] as well as unsupervised and supervised methods [6]. Data transformation mainly consists of data scaling and data type transformation. Data scaling aims to normalize the data variables so that they appear equally important in data analysis as far as quantity concerned. Commonly used scaling methods include the max-min normalization, Z-score normalization, and decimal point normalization [18]. Data type transformation is often needed before the data mining. For example, conventional association rule mining (ARM) algorithms, such as the Apriori and frequent-pattern growth algorithms, can only handle categorical data (e.g., High, Medium, and Low), while the majority of BAS variables are numeric. Hence, it is necessary to transform numeric data into categorical data prior to the use of conventional ARM algorithms. Popular methods for such data type transformation include the equal-frequency binning, equal-interval binning, and entropy-based discretization [18]. Data reduction aims to improve the computation efficiency through reducing data dimensions. BAS data is normally stored in such a format that each row represents an observation sampled at a specific time instant and each column represents the values of a variable in all observations. Sampling techniques, such as random sampling and stratified sampling, are commonly used for the reduction of the row number. The reduction of the column number, or the selection of variables of interests and significance, can be

done mainly in three ways. The first one is to select the variables of interests based on domain knowledge. The second one is to adopt data reconstruction methods, such as the principal component analysis in which the new low-dimensional variables are the linear combinations of the original high-dimensional variables [6, 20]. The third one is to use the heuristic methods, such as the step-wise forward selection and backward elimination methods, to select the variables most relevant to the problem concerned [18]. Visualization plays an essential role in the early stage of a mining process. It enables the users to have a straightforward understanding about the data. Visualization methods vary in their functionalities. For instance, box plots and histograms are efficient to display the data distribution; scatter plots provide a way to show correlations; run charts are useful to present time series data. However it is always a challenge to visualize high-dimensional data simultaneously.

2.2. Data partitioning

Data partitioning is necessary considering that most building services systems are highly dynamic and inter-correlated. The values of the variables and the relationships between variables may vary dramatically under different operating conditions. As a result, mining the entire BAS data simultaneously may result in significant knowledge loss. Partitioning the BAS data into several subsets of unique patterns according to their intrinsic characteristics and then mining the individual subsets helps to efficiently discover more meaningful knowledge. Since the distance of the data in a subset is remarkably reduced, or the similarity of the data is greatly improved, the knowledge being discovered is more reliable. However, this kind of data partitioning should mainly rely on the data intrinsic characteristics and involves less domain knowledge to take the advantage of DM in discovering underlying knowledge. How to capture the data intrinsic characteristics is a critical issue

and many methods can be adopted. The significance tests and clustering analysis are recommended to capture the intrinsic characteristics and partition the BAS data.

Significance testing

Significance testing or hypothesis testing is a method to examine two mutually exclusive hypotheses, i.e., the null hypothesis H_0 and the alternative hypothesis H_a [18]. The hypotheses are formulated with the aim of rejecting the null hypothesis. A level of significance (α) should be defined prior to the test. α is essentially the Type *I* error, which refers to the chance of incorrectly rejecting a true null hypothesis. Typically, the Type *I* error is set as 10%, 5%, or 1%, depending on the user's desired level of confidence (i.e., 1- α). In this study, due to the huge amount of data, a more stringent value, 1%, is selected. A test statistic score under the null hypothesis can be calculated using the sample data. The test statistic can then be converted to a probability value for decision-making. If the resulting probability is smaller than the predefined α, the null hypothesis (H_0) can be rejected. Otherwise, the test fails to reject the H_0. The method has been widely used to identify the effects of variables on data behaviors in the industries of marketing, medical and social science [21].

This study adopts the analysis of variance (ANOVA) method to investigate the effects of one or more qualitative variables on the quantitative outcomes concerned. The H_0 claims that the qualitative variables have little effect on the outcomes, while H_1 states the opposite. The basic idea is to partition the total variance of a quantitative outcome into two parts, i.e., variance within each qualitative value and variance between different qualitative values. From these two parts, the mean squares of errors and the mean squares of effects can be obtained. The test statistic is the ratio of the mean squares of effects to the mean squares of errors. The test statistic follows the *F*-distribution and

a probability value can be obtained accordingly. If the probability is smaller than α, then the H_0 is rejected. In other words, the qualitative variables may have significant effect on the outcomes.

Clustering analysis

Clustering analysis partitions the data into a number of clusters with the aim of maximizing the similarities of the observations in the same cluster while minimizing those between clusters. The similarity can be measured by various methods, such as the Euclidean distance and the Manhattan distance. The clustering results can be evaluated by either the internal validation methods (e.g., Davies-Bouldin index, Silhouette index, and Dunn index) or the external validation methods (e.g., purity, *F*-measure, and normalized mutual information) [6]. Five clustering algorithms, i.e., *k*-means, partitioning around medoids (PAM), hierarchical clustering, entropy weighting *k*-means (EWKM), and fuzzy c-means clustering, are selected as the candidate algorithms and their performances are compared in this study. The parameters of these algorithms are fine-tuned using the Dunn index, which integrates the inter-cluster dissimilarity and cluster diameters to evaluate the clustering results. A larger Dunn index indicates a better clustering result. A detailed discussion about these methods can be found in [6, 22].

2.3. Knowledge discovery

While the previous two phases prepare the data for mining, the knowledge discovery phase covers the actual mining process. A large number of DM techniques are available and new DM techniques are constantly emerging. The selection of DM techniques depends on the problems under consideration, data availability and the level of domain expertise. The knowledge discovered may be in the forms of clusters, decision trees, association rules, and etc., which are suitable for developing

predictive models, detecting and diagnosing abnormalities and developing optimization strategies. For example, the association rules and decision trees can be used for diagnostics. If the new observations violate the association rules, there is a high possibility that something abnormal occurred. Then, the decision trees can be used to find the source of the abnormality by deducing the variables which contribute the most to this kind of violation. Since building services systems are well understood nowadays, the domain knowledge about them is rich. Therefore, supervised DM techniques may not make significant contribution to the knowledge discovery. By contrast, unsupervised techniques are more capable of discovering unknown knowledge from the massive BAS data.

Association rule mining (ARM) is a popular unsupervised DM technique and it has been adopted in retail, marketing, and health care [23]. Compared with other forms of knowledge discovered by DM, interpretation of the association rules using domain knowledge is more convenient and utilization of the rules is more straightforward. Some efforts have been made on the application of ARM in the building field. Yu et al. [24] adopted the frequent-pattern growth algorithm to derive rules from the operational data of an air-conditioning system. The rules discovered were used to detect energy waste and component faults. Cabrera and Zareipour [25] presented the application of ARM in detecting lighting energy waste. The simulation results showed that up to 70% of energy use could be saved using the energy saving measures derived from the rules. Xiao and Fan [26] adopted the Apriori algorithm for BAS knowledge discovery. The association rules successfully identify non-typical and abnormal conditions in building operation. However, there are two major obstacles to applying ARM to the BAS data. Most of the conventional ARM algorithms, such as the Apriori

and FP-growth algorithms [6], can only handle categorical data, such as “High”, “Medium” and “Low”. However, almost all BAS data, such as power, temperature, humidity, flow rate and pressure, are numeric. Therefore, it is necessary to transform the numeric data to categorical data before using conventional ARM algorithms. In practice, it is very difficult to determine the intervals for the categories of “High”, “Medium” and “Low”, since BAS variables generally present large varieties. Secondly, ARM can usually generate a large number of rules. For instance, nearly 500 rules were derived in the work presented in [24] and [26]. Selecting useful rules is very challenging and time-consuming. This study adopts a new ARM technique, i.e. the Quantitative Association Rule Mining (QARM), to overcome the first obstacle. A novel rule selection method is also developed for fast selection of potentially useful rules in the post-mining phase.

Quantitative association rule mining (QARM)

The rule format for quantitative association rules is as follows: $\{A \in [a_1, a_2]\} \rightarrow \{B \in [b_1, b_2]\}$, where A and B are numeric variables, and a_1, a_2, b_1, b_2 specify the intervals for each numeric variable. $\{A \rightarrow B\}$ is called the rule pattern. In general, association rules are derived by defining two parameters, i.e., the minimum thresholds of support and confidence. The support of a rule is the joint probability of the antecedent and consequent, as defined in Equation (1). The confidence of a rule is the conditional probability of the consequent, given the antecedent, as shown in Equation (2). Only those rules meet the thresholds are derived and considered to be meaningful. The support threshold can be defined with great flexibility. A higher support threshold tends to find rules that happen more frequently, and vice versa. A low support threshold will lead to a dramatic increase in the number of association rules obtained, and consequently the post-mining will be time-consuming. The

confidence threshold should be maintained at a high level, e.g., above 85%, to ensure the association strength of the discovered rules. Another statistic, i.e., the lift, is commonly used to evaluate the "interestingness" of a rule. Lift is the ratio of the confidence to the support of the consequent, as defined by Equation (3). The lift can be considered as a measure of the dependence strength between the antecedent and the consequent. If the lift is larger than 1, it indicates that the occurrence of the antecedent positively affects the occurrence of the consequent, or if the probability of occurrence of the antecedent is high, the probability of occurrence of the consequent is also high. By contrast, a lift value smaller than 1 indicates a reversed correlation, which means, if the probability of occurrence of the antecedent is high, the probability of occurrence of the consequent is low. If the lift equals 1, it indicates that the antecedent and the consequent are independent and hence, the rule has little practical value. Generally speaking, the larger the lift value deviates from 1, the more interesting the rule is.

$$Support(A \rightarrow B) = P(A \ and \ B) \quad (1)$$

$$Confidence(A \rightarrow B) = P(B|A) = \frac{P(A \ and \ B)}{P(A)} \quad (2)$$

$$Lift(A \rightarrow B) = \frac{P(B|A)}{P(B)} = \frac{P(A \ and \ B)}{P(A)P(B)} \quad (3)$$

This study adopts the quantitative association rule mining (QARM) as the primary DM technique in the knowledge discovery phase. The QuantMiner [23, 27] is selected as the mining algorithm. The intervals (a_1, a_2) and (b_1, b_2) are determined by compromising the gain of an association rule and the length of the intervals. The gain of an association rule is defined by Equation [4], where *MinConf* is

the predefined minimum confidence threshold. The fitness function, which takes into account both the gain of the association rule and the length of the intervals, is defined by Equation [5]. Genetic algorithm is used to maximize the fitness function. Rules with large gains and small intervals are preferred [23, 27].

$$Gain(A \rightarrow B) = Support(A \cap B) - MinConf \times Support(A) \quad (4)$$

$$Fitness(A \rightarrow B) = Gain(A \rightarrow B) \times \coprod_{A_i \in A_{num}} \left[1 - \frac{size(I_{A_i})}{size(A_i)}\right]^2 \quad (5)$$

Where A_{num} refers to the number of numeric variables presented in the rule pattern$\{A \rightarrow B\}$; $I_{A_i}$is the interval of A_i; $size(A_i)$ is the range of A_i; $size(I_{A_i})$ is the length of the identified interval.

2.4. Post-mining

The post-mining phase performs three tasks, i.e., selection, interpretation, and utilization of the knowledge discovered. A novel approach is proposed to efficiently select potentially useful association rules in this study. Knowledge interpretation usually requires domain knowledge to explain the knowledge discovered by DM. Knowledge utilization aims to convert the knowledge into actionable measures for enhancing the building operational performance.

Rule selection

As above mentioned, the support and confidence are used to evaluate rules, and only those rules with the support and confidence meeting the predefined thresholds are considered. However, hundreds of rules may still be obtained, although the thresholds of the support and confidence are conservatively set. Selecting potentially useful rules by individual inspection is extremely time-consuming. It is noticed that the lift is helpful in selecting potentially useful rules. The larger the lift value deviates from 1, the more interesting the rule is. In this study, a novel rule selection approached based on the

lift is proposed for fast selection of potentially useful rules.

The massive BAS data are divided into several subsets according to data intrinsic characteristics in the 2nd phase and each subset will be mined separately in the 3rd phase. Similar rules with the same rule pattern may be obtained from mining different subsets. Such rules specify the associations between the same variables, but the intervals of the antecedents and consequents are different. These similar rules are of particular interest. If the lifts of these rules are more or less the same, the dependence strength between the antecedent and the consequent is consistent and stable under all operating conditions represented by corresponding subsets. If the lifts of these rules have large variations, the dependence strength of the association is influenced by the operating conditions, which is worthy of further investigation. The possible reasons for the large variations include the change of operating strategy and abnormalities occurred. In view of this, a rule selection approach is proposed for fast selection of potentially useful rules. The standard deviation of the lifts (SD-Lift) of the similar rules obtained from mining different subsets is calculated. Those rules, which result in a high SD-Lift, are then inspected individually to find the actual reasons causing the large lift variations.

Rule utilization

The knowledge discovered by DM can be used for various purposes, including prediction, diagnosis, and optimization. In this study, a method for utilizing the association rules for diagnosing abnormality in operation is proposed. All the rules obtained from mining one subset are utilized to build the knowledge base for the corresponding operating condition. Each new observation is examined against the rules in the corresponding knowledge base. A rule is violated if the observation

meets the antecedent but fails to meet the consequent. Since the lift value indicates the dependence strength between the antecedent and the consequent, the rules with larger lift values are more significant than those with smaller lift values. As a result, if an observation violates a rule with larger lift, the violation is more serious. Accordingly, an abnormality degree (AD) of an observation is proposed as shown in Equation (6), which measures the seriousness of the violation against all rules in the corresponding knowledge base. In Equation (6), *1* is subtracted from the lift values, as a lift value of 1 indicates independence between the antecedent and the consequent. A lift value smaller than 1 means the probability of occurrence of the consequent is low when the probability of occurrence of the antecedent is large. Therefore, if a new observation violates a rule with a lift value smaller than 1, it is actually normal, rather than abnormal, and the AD should be decreased.

$$AD = \sum_{i=1}^{n} (lift_i - 1) \qquad (6)$$

Where n is the number of rules being violated, and $lift_i$ is the lift value of the i^{th} rule being violated.

3. Implementation of the DM-based mining framework

3.1. Description of the raw BAS data

The BAS data were retrieved from the tallest building in Hong Kong, i.e. the International Commerce Centre (ICC). The building is served by a central chilling system consisting of six identical high-voltage centrifugal chillers. Each chiller is associated with a constant-speed primary chilled water pump and a constant-speed condenser water pump. The primary-secondary chilled water loop is used to transfer the cooling energy to the demand side. The heat dissipated from the chiller condensers is rejected by 11 evaporative cooling towers. A more detailed description of the system can be found in [28]. ICC is equipped with an advanced BAS system. The central chilling

system alone is monitored and controlled with more than 500 measurements. ICC was awarded with the Intelligent Building of 2011 by the Asian Institute of Intelligent Buildings.

The power consumptions of almost all electrical equipment, including the chillers, pumps, fans, vertical transportation system and lighting system are recorded. One-year BAS data (from January 2013 to December 2013) with a sampling interval of 15-minute are analyzed in this study. The whole data set includes 29,757 observations and each observation records 158 variables. The variables include the date and time (i.e., year, month, day, hour, minute, day type), power consumptions of 12 sub-systems including water-cooled chillers (WCC), cooling towers (CT), primary chilled water pumps (PCHWP), secondary chilled water pumps (SCHWP), condenser water pumps (CDWP), air-handling units (AHU), primary air-handling units (PAU), normal power and lighting (NLTG), essential power and lighting (ELTG), vertical transportation system (VTS), plumbing and drainage (PD) and mechanical ventilation system (MV), as well as and the measurements of temperature, flow rate, pressure and etc. in the waterside system of the air conditioning system.

The BAS data contains significant amount of missing values, dead values and outliers. The proportion of the missing values is around 1.28% in the ICC BAS data and they are filled up using a simple moving average method with a window size of 10 in this study. When the values of a variable missed for longer than 2 hours, the corresponding observations are discarded, since the operating conditions and system behaviors may significantly change during that period. The outliers whose values are obviously outside the normal range can be easily detected using domain expertise. For instance, the outdoor temperature measurements higher than 50^{o}C or lower than 0^{o}C are discarded because it never happened in Hong Kong. Dead values refer to those measurements whose values

remain the same for a rather long period. In this study, all the observations suffering from the "dead values" for longer than 2 hours are discarded.

3.2 Identification of typical building operating patterns

The 2nd phase in the framework aims to identify typical building operating patterns so as to partition the original massive data into several subsets. Building operation is mainly influenced by climate conditions and occupancy level. Moreover, people are very much concerned about building energy efficiency and indoor environment quality (IEQ). Identification of building operating patterns related to energy consumption and IEQ can help to explore means to enhance them. IEQ can be assessed by monitoring the concentrations of the indoor CO_2 and other typical pollutants, indoor illuminance and noise levels, and etc. However, such measurements are usually not available in today's BASs, including the BAS of ICC. The power consumptions of various components are well recorded in ICC. Therefore, this study focuses on typical power consumption patterns.

The identification process is undertaken in two steps. Firstly, ANOVA is applied to analyze the significance of time variables (i.e., month, date, hour, minute, day) to the aggregated power consumption. Then, clustering analysis is used to find the optimal number of clusters which the original data can be partitioned into according to the significant variables, as well as to determine the cluster membership of each observation.

In this study, the level of Type *I* error, α, is defined as 1%. The ANOVA results were shown in Table-1. It indicates that only three time variables, i.e., month, day, and hour, result in a probability smaller than the specified Type *I* error. Therefore, these three variables have significant effects on the aggregated building power consumption. Five clustering analysis methods (i.e., k-means,

hierarchical clustering, PAM, fuzzy c-means, and EWKM) are adopted to partition the large BAS data. The clustering analysis is performed in the sequence of “month”, “day”, and “hour” to avoid the conflict of cluster memberships. To determine the cluster membership in terms of the variable “month”, all the power consumptions of the 12 sub-systems are scaled using max-min normalization. Then, the mean and standard deviations of the power consumption of each sub-system are calculated for each month, resulting in a feature data set with 24 variables. Clustering analysis is performed based on the feature data set. When determining the cluster membership in terms of the variable “day”, the original observations in the months which are grouped in the same cluster are analyzed together. Features are then calculated for each day (i.e. Monday to Sunday) and clustering analysis is applied to find the cluster membership. Similar approach is adopted in determining the cluster membership in terms of “hour”.

Table-1 ANOVA testing results

Variable	DOF	Sum of squares	Mean sum of squares	F-test statistics	Probability
Month	11	34,233,661	3,112,151	759.46	< 1%
Date	30	1,361,300	453,800	0.37	78%
Hour	23	220,948,230	9,606,445	2,344.25	< 1%
Minute	3	2,401	800	0.20	90%
Day	6	89,532,444	14,922,074	3,641.42	< 1%

Fig. 2 and 3 illustrate the clustering results in terms of “hour”. Fig. 2 presents the Dunn indices of different clustering algorithms and cluster numbers. The maximum Dunn index can be obtained when EWKM is used and the cluster number is 2. Therefore, such combination is selected to perform

the clustering analysis. Two parameters of EWKM, i.e., the weight distribution parameter (λ) and the convergence threshold (δ), are determined using the Dunn index and they are 0.15 and 0.0001, respectively. Fig. 3 illustrates the cluster membership in terms of the time variable "hour". It can be found that the majority of observations collected during 8:00 to 20:00 are grouped in the 1st cluster and the rest observations are grouped in the 2nd cluster. In Hong Kong, 8:00 to 20:00 are normally the office hours and the other hours are non-office hours. The power consumptions during office hours and non-office hours are very different due to the different operating conditions, particularly occupancy levels. Therefore, the clustering result is consistent with the domain knowledge.

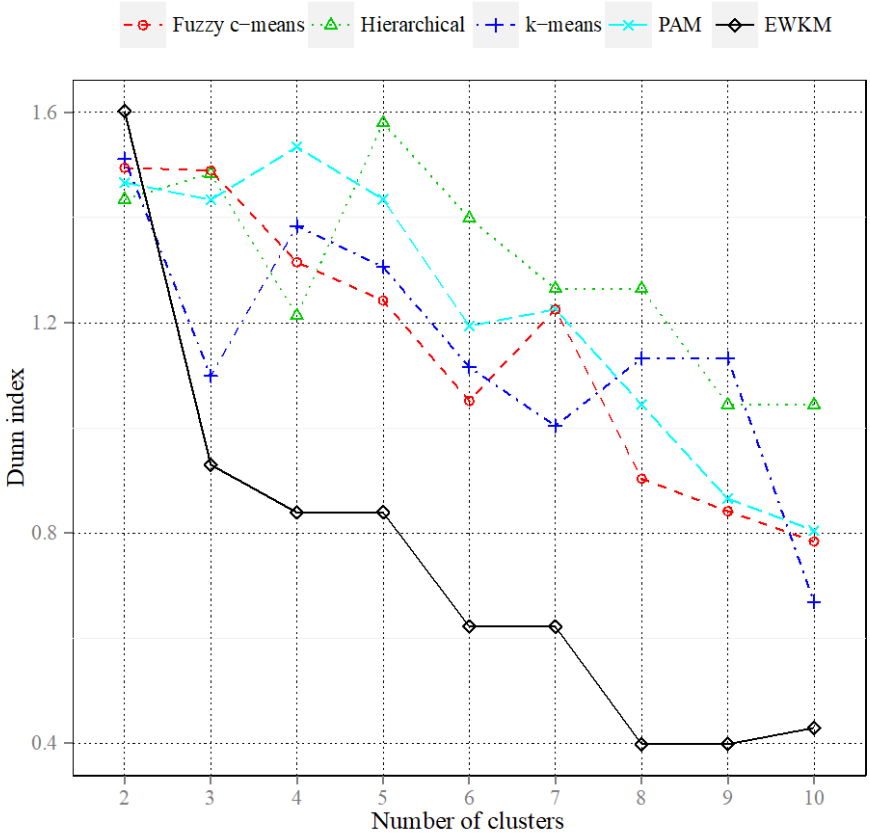

Fig. 2 Comparison of clustering algorithms for clustering in terms of "hour"

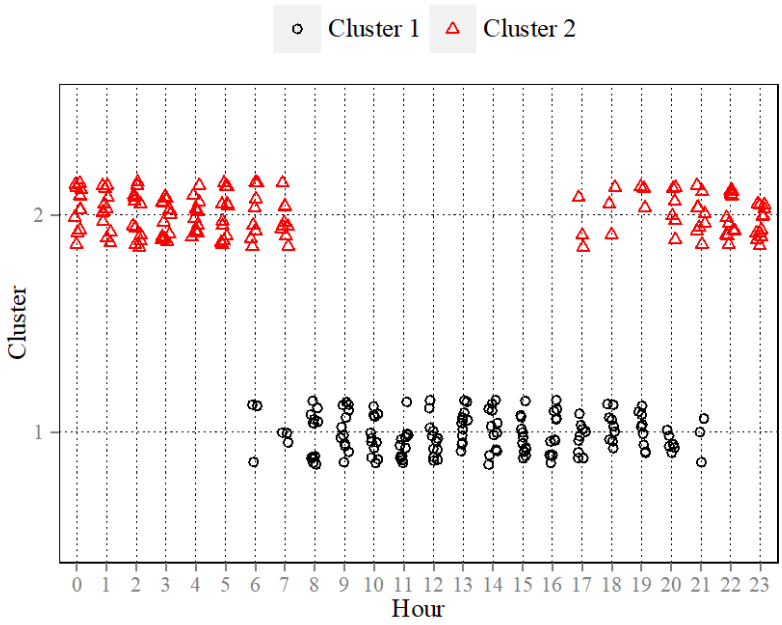

Fig. 3 Cluster membership in terms of “hour”

Table-2 summarizes the overall clustering results. The optimal cluster number is 2 for all cases. EWKM is chosen as the clustering algorithm in terms of “month” and “hour”, while the hierarchical method is chosen for “day”. When the data is grouped in terms of “month”, data collected from June to October are grouped in one cluster, and the data from the other months are grouped in the second cluster. June to October are normally the hot season in Hong Kong with the higher outdoor temperature and relative humidity. Cooling demand in hot season is very large. By contrast, the other months are cool season and cooling demand is relatively low. The two clusters in terms of “day” are corresponding to weekdays (i.e. Monday to Friday) and weekends (i.e. Saturday and Sunday), respectively. As a result, the original large BAS data are partitioned into eight clusters or subsets, and each cluster is defined by a combination of the three time variables as shown in Table 3. The clustering results are reasonable, since each cluster has its unique power consumption pattern considering the climate conditions and occupancy levels.

Table-2 Summary of the clustering results

Clustering Variables	Clustering Algorithms	Optimal Cluster Number	Cluster Membership
Month	EWKM	2	{6-10 }; {1-5 & 11-12}
Day	Hierarchical	2	{Monday to Friday}; {Saturday and Sunday}
Hour	EWKM	2	{8:00-20:00}; {0:00-8:00 & 20:00-0:00}

Table-3 Summary of the eight clusters

Cluster	Month type	Day type	Hour type
1	Hot season	Weekdays	Office hours
2	Hot season	Weekdays	Non-office hours
3	Hot season	Weekends	Office hours
4	Hot season	Weekends	Non-office hours
5	Cool season	Weekdays	Office hours
6	Cool season	Weekdays	Non-office hours
7	Cool season	Weekends	Office hours
8	Cool season	Weekends	Non-office hours

3.3. Knowledge discovery using QARM

QuantMiner is selected to mine the eight data subsets separately. The minimum thresholds of the support and confidence are set as 0.1 and 0.9. The support threshold is set at a relatively low level and the confidence is relatively high, with the aim of discovering those associations that are not necessarily very frequent but very strong. Although QuantMiner is capable of mining association rules with multiple variables in both antecedents and consequents, this study only focuses on the association rules with only one variable in the antecedent and consequent respectively for easy interpretation.

Each of the eight subsets generates 534 rules and 4,272 rules are obtained in total. The post-mining

method described in section 2.4 is adopted for rule selection, interpretation, and utilization. The SD-Lifts of similar rule patterns obtained from mining different subsets are calculated. It is found that the majority of the rules with the same rule patterns have a SD-Lift smaller than 0.2. The rule patterns having a large SD-Lift have been selected for further analysis. An example is presented in the following section.

4. Applications of the knowledge discovered

4.1. Identification of the change in building operating strategies

One rule pattern $\{WCC \rightarrow PAU\}$ draws special attention as it has a large SD-Lift of 0.3. $\{WCC \rightarrow PAU\}$ describe the associations between the chiller power consumption (WCC) and the PAU fan power consumption (PAU). The details of the rules with this rule pattern are shown in Table-4. The clusters are numbered in accordance to Table-3. The rules obtained from mining Cluster 1 and 8 are interpreted with domain knowledge here. Cluster 1 is corresponding to “Hot season”, “Weekdays” and “Office hours”, which means hot climate and high occupancy level. Cluster 8 is corresponding to “Cool season”, “Weekends” and “Non-office hours”, which means cool climate and low occupancy level. According to domain knowledge, the demand for cooling is higher under hot climate, and the demand for outdoor air ventilation is higher for high occupancy level. Therefore, both WCC and PAU of Cluster 1 should be higher than those in Cluster 8, which can be seen from the intervals of WCC and PAU of Cluster 1 and Cluster 8 in Table 3. Meanwhile, if the “day type” and “hour type” are the same which means the occupancy level are similar, the WCC intervals in “Hot season” should be higher than those of “Cool season”. Rules obtained from Cluster 1 and Cluster 5 in Table 3 also support this argument. However, it is observed that when the “day type” and “hour type” are the

same, the upper limits of the PAU fan power consumption in the hot season are much lower than those in the cool season, even though the lower limits are similar. Taking the rules obtained from Cluster 1 and Cluster 5 as example, the lower limits of PAU are 330.4 kW and 328.7 kW, which are quite close. However, the upper limits, 416.2 kW and 462.8 kW, are quite different. Similar phenomenon can be observed for Cluster 2 and 6, Cluster 3 and 7, as well as Cluster 4 and 8. This phenomenon disobeys the domain knowledge which tells that the PAU fan power consumption should be similar for the same "day type" and "hour type". Further investigation is carried out to exploit the root cause.

Table-4 Summary of the rule pattern $\{WCC \rightarrow PAU\}$

Cluster	WCC	PAU	Supp.	Conf.	Lift
1	[2602.3, 3133.7]	[330.4, 416.2]	0.35	0.95	1.19
2	[884.3, 963.0]	[101.9, 210.3]	0.27	0.98	1.29
3	[740.3, 1603.7]	[98.5, 221.1]	0.36	0.97	1.31
4	[763.3, 906.4]	[94.3, 173.3]	0.37	0.97	1.07
5	[1797.7, 2444.1]	[328.7, 462.8]	0.27	0.99	1.74
6	[718.3, 811.7]	[94.7, 270.7]	0.26	0.96	1.64
7	[932.5, 1115.6]	[94.0, 324.7]	0.28	0.99	1.92
8	[679.0, 791.4]	[92.9, 218.2]	0.35	0.98	1.65

A decision tree is developed using the CART algorithm [6] to explore the underlying relationship among time variables and the PAU fan power consumption. The PAU fan power consumption is the output, while the "month", "day" and "hour" are selected as inputs. For easy interpretation, the tree depth is limited to 2, which means the remotest terminal node can be reached from the root node through 2 splits. As shown in Fig. 4, four terminal nodes (i.e., Node 3, 4, 6, and 7) are derived to represent the four levels of PAU power consumption. The associated boxplots show the distribution

of the PAU power consumption at each terminal node. The algorithm selects the “hour” as the splitting variables at the root node (i.e., Node 1). It automatically divides the “hour” into two groups, i.e., non-office hours {0, 1, 2, 3, 4, 5, 6, 7, 21, 22, 23} and office hours {8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20}, which is in accordance with the cluster membership discovered by clustering analysis. Similarly, Node 5 selects the “day” for splitting and the results are the same as that obtained from the clustering analysis. Node 2 uses the “month” for splitting; however, the grouping of months (June to December in one group, and January to May in the other group) is different from the clustering results (June to October in one cluster, and the other months in the other cluster).

The right side of the tree states that when the observations are measured during the office hours, the PAU power consumption is closely related to the “day type”. It is observed that the PAU power consumption in weekdays is significantly higher than that in weekends. This is reasonable because people normally don’t work in weekends which results a large drop in the occupancy level. The left side of the tree states that when the observations are recorded during the non-office hours, the PAU power consumption is closely related to the “month”. It is noted that the PAU power consumption during the first five months (i.e., Jan to May) is higher than that during June to December, which cannot be explained by domain knowledge.

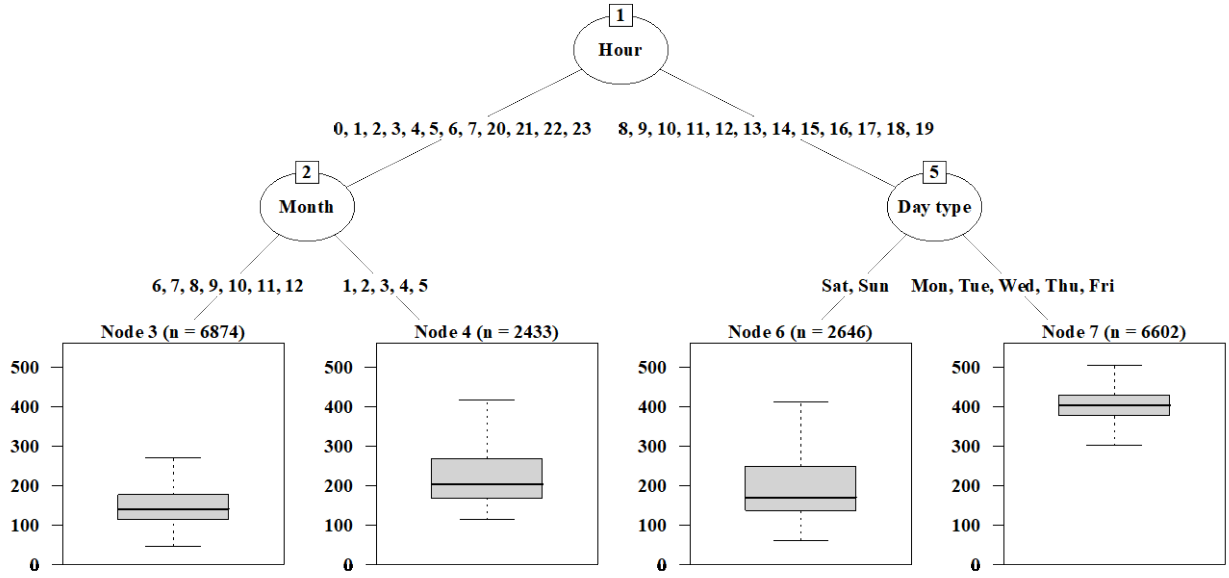

Fig. 4 Decision tree for PAU power consumption

After consulting the operation staff, it is found out that the PAU operating strategy did change in June 2013. Before the change, the PAU fan speed was controlled at three levels, i.e., 0 L/s, 960 L/s, and 1200 L/s. If the CO_2 concentration is below 800 ppm, 960 L/s is used; otherwise, 1200 L/s is used. Starting from June 2013, the demand-controlled ventilation (DCV) strategy is implemented. Under this strategy, the fresh air flow rate is continuously controlled between 850 L/s and 1200 L/s to maintain indoor CO_2 concentration at its set-point. That's why the tree model adopts the "month" as the splitting variable at Node 2, which also indicates that the DCV strategy results in more energy saving during non-office hours. This is reasonable because that the occupancy level during non-office hours is quite low and the demand for outdoor air ventilation is also low. The energy saving during the office hours is not that obvious because the occupancy level is quite stable throughout the year.

4.2. Identification of non-typical building operating conditions

A number of continuous observations in Cluster 5 (i.e., Cool season, Weekdays, Office hour) are

found to have high abnormality degrees (ADs). The examples of the rules being violated are summarized in Table-5. “Temp_rtn_ch” and “Temp_sup_ch” refer to the return and the supply chilled water temperature respectively. The first two rules state that the NLTG power consumption has associations with the WCC power consumption and the return chilled water temperature. The lower limits of NLTG are approximately 500 kW. However, the actual NLTG measurements of the observations with high ADs are around 430 kW. The 3rd and the 4th rules describe the relationship among PAU fan power consumption, the supply chilled water temperature, and WCC. The lower limits of PAU are around 300 kW, while the PAU fan power consumptions in the observations with high ADs are around 250 kW. The last two rules describe the associations among VTS, SCHWP, and WCC. The VTS in the observations with high ADs are smaller than the lower limits specified in each rule.

Further investigation shows that all these observations with high ADs are from Wednesday May 1, 2013, which is a public holiday in Hong Kong. The profiles of the NLTG, PAU and VTS power consumption on May 1, 2013 are shown in Figs. 5. The corresponding average power consumptions in Cluster 5 are also plotted for comparison. It is obvious that the measurements on May 1, 2013 are much lower than the corresponding average values. This case shows that the abnormality degree is effective in diagnosing non-typical and abnormal building operations.

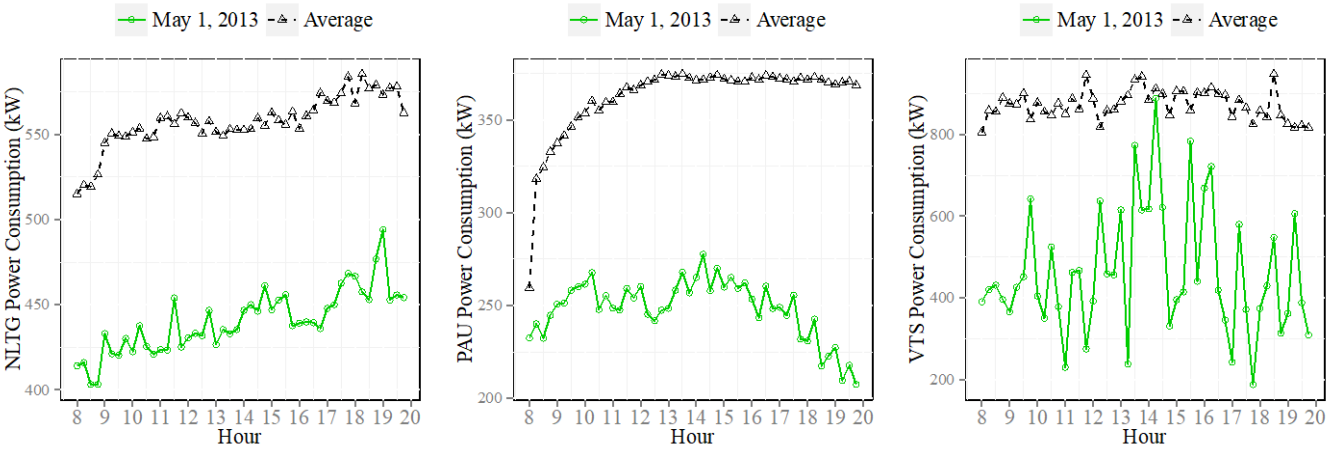

Fig. 5 NLTG, PAU and VTS measurements on May 1, 2013 against the average values

Table-5 Summary of rules being violated

No.	Antecedent	Consequent	Supp.	Conf.	Lift
1	WCC in [1576.8, 2438.2]	NLTG in [510.3, 688.7]	0.30	0.96	1.66
2	Temp_rtn_ch in [9.8, 10.6]	NLTG in [506.1, 659.6]	0.33	0.97	1.47
3	Temp_sup_ch in [6.4, 6.9]	PAU in [303.5, 477.8]	0.27	0.97	1.61
4	WCC in [1797.7, 2444.1]	PAU in [318.7, 422.8]	0.17	0.99	1.64
5	SCHWP in [73.1, 109.4]	VTS in [401.3, 1461.1]	0.30	0.95	1.34
6	WCC in [1147.4, 1669.8]	VTS in [418.8, 1580.5]	0.28	0.97	1.38

Furthermore, it is found that large ADs take place during the similar periods every day. Fig. 6 shows the profiles of the means of ADs on Friday, Saturday, and Sunday. The profiles on weekdays are very similar, so only the profile on Friday is shown here as an example. There are two obvious spikes on Friday and other weekdays as well, which take place during 6 a.m. to 9 a.m., and 7 p.m. to 9 p.m. Three main spikes are observed on Saturdays, and they are recorded during 7 a.m. to 9 a.m., 2 p.m. to 4 p.m., and 7 p.m. to 9 p.m. The profile on Sundays is relatively flat, and only one small spike is observed between 7 a.m. and 9 a.m. The results are in accordance with the domain expertise. The office hours for typical office buildings in Hong Kong are from 8:00 to 20:00 in weekdays, and

8:00 to 14:00 on Saturdays. The building system performs either a stage-up or a stage-down process during these periods. The transient changes normally result in very different operation behaviors. For instance, during the stage-up process, chilled water pumps usually consume much more power due to the motor starting characteristics. Fig. 6 shows that a typical stage-up or stage-down process normally last for around 2 hours.

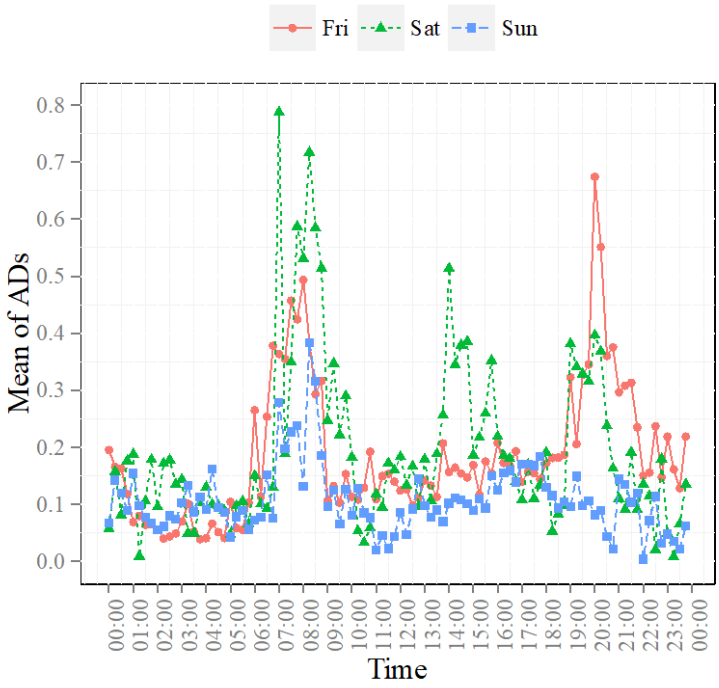

Fig. 6 AD means on Fridays, Saturdays, and Sundays

4.3. Fault detection of power consumption sensors

It was found that the rules related to the VTS are frequently violated during the period between Mar 17, 2013 and Apr 22, 2013. Table-6 presents three examples of the rules being violated, which describe the associations between VTS and three main HVAC subsystems, i.e. WCC, PAU and SCHWP.

Table-6 Examples of the rules being violated related to VTS

No.	Antecedent (kW)	Consequent (kW)	Supp.	Conf.	Lift

1	WCC in [817.0, 1403.2]	VTS in [490.3, 1442.4]	0.29	0.96	1.48
2	PAU in [371.5, 406.9]	VTS in [565.5, 1614.5]	0.37	0.97	1.37
3	SCHWP in [60.4, 83.1]	VTS in [537.5, 1556.5]	0.31	0.96	1.36

The VTS power consumptions in the observations violating the rules are smaller than the lower limits of VTS in the rules. Fig. 7 shows the VTS power consumption of the abnormal observations against those of the normal observations. It is shown the VTS power consumptions of abnormal observations are much lower. The power consumption of VTS in ICC consists of five parts, i.e., lifts in the car parking area, office shuttle lifts, office service lifts, fireman lifts, and escalators. Further investigation shows that during the above-mentioned period, the power meter for the office service lifts broke down and no value was recorded. Therefore, the aggregated power consumption for the VTS was smaller.

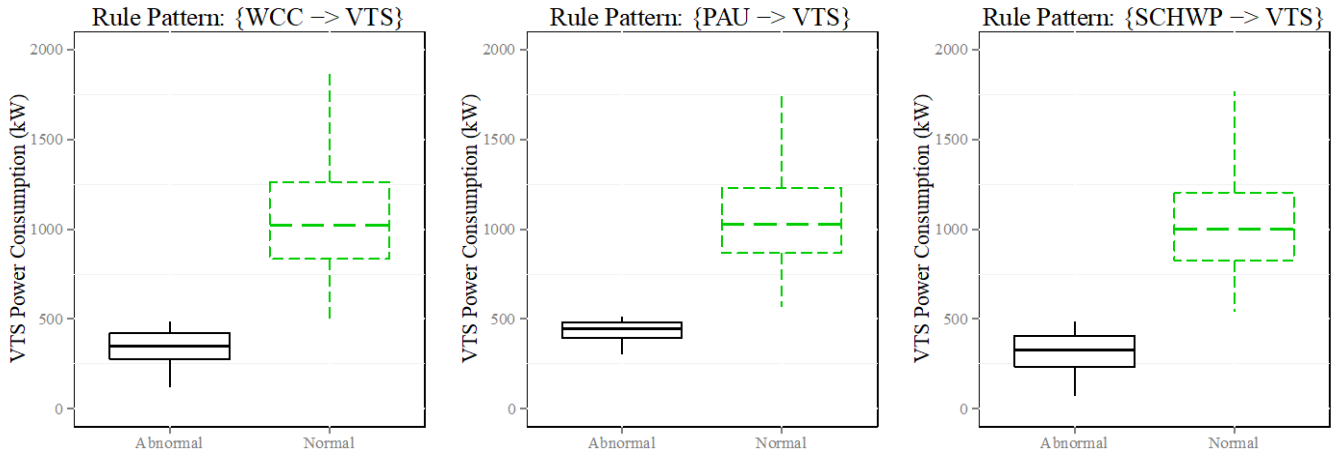

Fig. 7 VTS power consumptions of normal and abnormal observations

5. Conclusions

The building automation industry is eager for a powerful tool for analyzing the massive data stored in the building automation system. As a promising technology for handling big data, data mining (DM) provides both opportunities and challenges for analyzing the massive BAS data. The

challenges are mainly caused by the knowledge gap between building professionals and DM experts. A generic DM-based framework is proposed in this study to build a bridge between the advanced DM techniques and the massive BAS data. Considering the rich domain knowledge in the building field, unsupervised DM techniques are recommended as the primary means of discovering underlying data structures and relationships in BAS data. The framework is also specifically designed to address the issues of low quality and complexity of BAS data. It is recommended that the entire data mining process is conducted in four phases. Typical methods for each phase are briefly introduced. The framework provides a reference for developing DM-based tools for knowledge discovery in massive BAS data and application of the knowledge discovery for building diagnostics.

The framework has been implemented in analyzing the BAS data of International Commerce Centre, the tallest commercial building in Hong Kong. For the first time in the building field, the advanced quantitative association rule mining (QARM) is adopted to discover the association rules in BAS data. QARM overcomes the weakness of conventional ARM in mining numeric data like most of BAS data. Two indices of high practical values are defined in this study to facilitate the post-mining process, i.e. the standard deviation of lift (SD-Lift) of rules with similar rule pattern and the abnormality degree (AD). SD-Lift can help to fast select useful rules from a large number of rules obtained in ARM, which is a major obstacle to the application of ARM. AD provides a generic method of using the association rules for detecting abnormalities. These two indices are proven to be valuable for applying the knowledge discovered by DM (i.e. association rules in this case) to building diagnostics. The change of operation strategy, non-typical and abnormal operations and

sensor fault occurring during operation in ICC are successfully detected and diagnosed. The open-source software *R* was used to perform all the DM techniques used in this study.

The framework developed in this paper serves as the prototype of a more comprehensive and sophisticated DM-based solution for discovering and applying knowledge hidden in the massive BAS data. Besides clustering analysis, ARM and decision tree, more advanced DM techniques will be investigated in future. Effort will made on developing general domain knowledge representation methods (like SD-Lift and AD proposed in this study) for translating the knowledge discovered by DM (e.g., correlations, clusters, association rules, patterns decision trees, and models) to meaningful and actionable knowledge.

Acknowledgements

The authors gratefully acknowledge the support of this research by the Research Grants Council (RGC) of the Hong Kong SAR (152181/14E).

References

[1] Waide, P., Ure, J., Karagianni, N., Smith, G., Bordass, B., The scope for energy and CO_2 savings in the EU through the use of building automation technology, Final Report for the European Copper Institute, August 10, 2013.

[2] International Energy Agency (IEA), accessed on Apr 9, 2014, https://www.iea.org/aboutus/faqs/energyefficiency/

[3] Machairas, V., Tsangrassoulis, A., Axarli, K., Algorithms for optimization of building design: A

review, Renewable and Sustainable Energy Reviews 31 (C) (2014) 101-112.

[4] Katipamula, S., Brambley, M.R., Methods for fault detection, diagnostics, and prognostics for building systems-A review, Part I & II, HVAC&R Research 11 (1-2) (2005) 3-25, 169-187.

[5] 10 breakthrough technologies, MIT Technology Review, January/February 2001, Massachusetts Institute of Technology, Cambridge, MA, USA.

[6] Tan, P.N., Steinbach, M., Kumar, V., Introduction to data mining, 1st edition, 2005, Addison-Wesley Longman Publishing, Boston, MA, USA.

[7] Amin-Naseri, M.R., Soroush, A.R., Combined use of unsupervised and supervised learning for daily peak load forecasting, Energy Conversion and Management 49 (6) (2008) 1302-1308.

[8] Dong, B., Cao, C., Lee, S.E., Applying support vector machines to predict building energy consumption in tropical region, Energy and Buildings 37 (5) (2005) 545-553.

[9] Chou, J.S., Hsu, Y.C., Lin, L.T., Smart meter monitoring and data mining techniques for predicting refrigeration system performance, Expert Systems with Applications 41 (5) (2014) 2144-2156.

[10] Khan, I., Capozzoli, A., Corgnati, S.P., Cerquitelli, T., Fault detection analysis of building energy consumption using data mining techniques, Energy Procedia 42 (57) (2013) 557-566.

[11] Hou, Z.J., Lian, Z.W., Yao, Y., Yuan, X.J., Data mining based sensor fault diagnosis and validation for building air conditioning system, Energy Conversion and Management 47 (15-16) (2006) 2479-2490.

[12] Magoules F., Zhao, H.X., Elizondo, D., Development of an RDP neural network for building energy consumption fault detection and diagnosis, Energy and Buildings 62 (18) (2013) 133-138.

[13] Kusiak, A., Li, M.Y., Tang, F., Modeling and optimization of HVAC energy consumption, Applied Energy 87 (10) (2010) 3092-3102.

[14] Kusiak, A., Tang, F., Xu, G.L., Multi-objective optimization of HVAC system with an evolutionary computation algorithm, Energy 36 (5) (2011) 2440-2449.

[15] Ahmed, A., Korres, N.E., Ploennigs, J., Elhadi, H., Menzel, K., Mining building performance data for energy-efficient operation, Advanced Engineering Informatics, 25 (2) (2011) 341-354.

[16] Chang, Y.C., Sequencing of chillers by estimating chiller power consumption using artificial neural networks, Building and Environment 42 (1) (2007) 180-188.

[17] Zhang, S.C., Zhang, C.Q., Yang, Q., Data preparation for data mining, Applied Artificial Intelligence 17 (5-6) (2003) 375-381.

[18] Hastie, T., Tibshirani, R., Friedman, J., The elements of statistical learning: Data mining, inference and prediction, 2nd edition, Springer Series in Statistics, New York, USA, 2009.

[19] Huang, G.S., Wang, S.W., Xiao, F., Sun, Y.J., A data fusion scheme for building automation systems of building central chilling plants, Automation in Construction 18 (3) (2009) 302-309.

[20] Xiao, F., Wang, S.W., Zhang, J.P., A diagnostic tool for online sensor health monitoring in air-conditioning systems, Automation in Construction 15 (4) (2006) 489-503.

[21] Larsen, R.J., Marx, M.L., Introduction to mathematical statistics and its applications, 4th edition, Pearson Prentice Hall, Saddle River, New Jersey, USA, 2006.

[22] Jing, L.P., Ng, M.K., Huang, J.Z., An entropy weighting *k*-means algorithm for subspace clustering of high-dimensional sparse data, IEEE Transactions on Knowledge and Data Engineering 19 (8) (2007) 1026-1041.

[23] Salleb-Aouissi, A., Vrain, C., Nortet, C., QuantMiner: A genetic algorithm for mining quantitative association rules, In the Proceedings of the 20th International Conference on Artificial Intelligence IJCAI, 2007, 1035-1040, Hyderabad, India.

[24] Yu, Z., Haghighat, F., Fung, C.M., Zhou, L., A novel methodology for knowledge discovery through mining associations between building operational data, Energy and Buildings 47 (50) (2012) 430-440.

[25] Cabrera, D.F.M., Zareipour, H., Data association mining for identifying lighting energy waste patterns in educational institutes, Energy and Buildings 62 (26) (2013) 210-216.

[26] Xiao, F., Fan, C., Data mining in building automation system for improving building operational performance, Energy and Buildings 75 (11) (2014) 109-118.

[27] Salleb-Aouissi, A., Vrain, C., Nortet, C., Kong, X.R., Rathod, V., Cassard, D., QuantMiner for mining quantitative association rules, Journal of Machine Learning Research 14 (1) (2013) 3153-3157.

[28] Ma, Z.J., Wang, S.W., Xu, X.H., Xiao, F., A supervisory control strategy for building cooling water systems for practical and real time applications, Energy Conversion and Management 49 (2008) (8) 2324-2336.