Applied Energy 253 (2019) 113492

Contents lists available at ScienceDirect

Applied Energy

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

An improved association rule mining-based method for revealing operational problems of building heating, ventilation and air conditioning (HVAC) systems

Chaobo Zhang[a], Xue Xue[b], Yang Zhao[a,*], Xuejun Zhang[a], Tingting Li[a]

[a] Institute of Refrigeration and Cryogenics, Zhejiang University, Hangzhou 310027, China
[b] Global Application Product Development Headquarters, Hanergy Thin Film Power Group, Shenzhen 518000, China

HIGHLIGHTS

- An improved association rule mining-based method is proposed for buildings.
- A kernel density estimation-based approach is applied for data preprocessing.
- An association rule comparison-based approach is proposed for post mining.
- This method can detect operational problems of HVAC systems effectively.
- This method can filter out about half of useless association rules effectively.

ARTICLE INFO

Keywords:
Data mining
Kernel density estimation
Association rule mining
Building operational performance
Building energy efficiency
Heating, ventilation and air conditioning systems

ABSTRACT

Energy wastes in heating, ventilation and air conditioning (HVAC) systems of buildings are very common due to lots of operational problems. It is in great need to develop data mining-based methods to discover these operational problems from the historical data of HVAC systems. In the past years, researchers had realized that association rule mining was one of the most effective algorithms to solve this problem. But, most of the mined operational patterns are useless. It is time-consuming to check them manually. In this study, an improved association rule mining-based method is proposed to enhance the performance of data mining and to filter out useless rules automatically. It contains three steps, i.e., data preprocessing, association rule mining and post mining. In the step of data preprocessing, a kernel density estimation-based approach is developed to filter out outliers automatically. And, a kernel density estimation-based approach is developed to transform numerical data into categorical data automatically. In the step of association rule mining, the FP-growth algorithm is utilized to extract raw association rules from the preprocessed data. In the step of post mining, a novel comparison-based approach is developed to reduce the amount of useless association rules. Evaluations are made using the historical operational data of the chiller plant of a commercial building. Results show that the proposed data preprocessing approaches are effective in outlier identification and data transformation. And, the proposed comparison-based approach can filter out 54.98% of the mined association rules automatically which are useless for discovering operational problems.

1. Introduction

The building sector has become the largest energy consumer in the world [1]. It contributes to more than one-third of the total global final energy consumption [2]. Heating, ventilation and air conditioning (HVAC) systems are the main energy consumers in public buildings which contribute to about 30–40% of buildings energy consumption [3]. However, HVAC systems are usually energy inefficient in practice [4,5]. About 15%-30% of energy used in HVAC systems is wasted for the reasons of sensor faults, device faults, performance degradations, improper control strategies, and so on [6]. Therefore, it is very valuable to discover these operational problems in time. Massive amounts of historical data of HVAC systems are available now with the popularity of building automation systems. It is possible to introduce artificial intelligence technologies such as data mining to discover operational problems from the historical data.

* Corresponding author.
E-mail address: youngzhao@zju.edu.cn (Y. Zhao).

https://doi.org/10.1016/j.apenergy.2019.113492
Received 24 December 2018; Received in revised form 5 June 2019; Accepted 7 July 2019

Nomenclature

f	density function
$x_1, x_2, \ldots, x_n$	independent and identically distributed measurements
K	kernel function
h	bandwidth
δ	threshold
α	factor of outlier threshold
m	threshold of the number of categories
C	consistency indicator
p	peak density
v	valley density
A	antecedent
B	consequent
support($A \rightarrow B$)	support of the association rule "$A \rightarrow B$"
confidence($A \rightarrow B$)	confidence of the association rule "$A \rightarrow B$"
lift($A \rightarrow B$)	lift of the association rule "$A \rightarrow B$"
$P(A \cup B)$	probability that the A and the B coincide
$P(B\|A)$	conditional probability of the B given the A
$P(A)$	probability that the A appears
$P(B)$	probability that the B appears
z'	normalized interval boundary
z	original interval boundary
D	Euclidean distance
l	lower interval boundary
u	upper interval boundary
β	factor of distance threshold
T	temperature (°C)
RH	relative humidity (%)
F	frequency (Hz)
P	power (kW)

Subscript

max	maximum
min	minimum
left	left-side
right	right-side
outlier	outlier
dist	distance
OUT	outdoor
CH	chiller
HP	heat pump
COWP	cooling water pump
CHWP	chilled water pump
SP	secondary pump
CT	cooling tower
OTCT	open-type cooling tower
CTCT	closed-type cooling tower
HE	heat exchanger
WDH	water distribution header
WCH	water collection header
SCHW	supply chilled water
RCHW	return chilled water
SG	supply glycol
RG	return glycol
TSG	total supply glycol
TRG	total return glycol
TSCOW	total supply cooling water
TRCOW	total return cooling water
TSCHW	total supply chilled water
TRCHW	total return chilled water
SB	skirt building
TB1-6	1–6 floors of the tower building
TB7-38	7–38 floors of the tower building

Data mining is an interdisciplinary subject with a goal to discover information from massive amounts of data using intelligent methods and transform the information into a comprehensible structure [7]. It has been widely applied in various fields. In the finance field, Pérez-Martín et al. adopted data mining technologies to analyze the data of home equity loans [8]. Chen et al. discovered customer behaviors from the data of retail marketing using data mining technologies [9]. In the medical treatment field, Wang et al. utilized data mining technologies to understand the data of healthcare [10]. Data mining technologies also had been widely utilized in the field of energy management. Torregrossa et al. [11] and Zhang et al. [12] both used data mining technologies to optimize the operation of waste water treatment systems. Heng et al. adopted data mining technologies with an aim of accurate solar radiation forecasting [13]. Astolfi et al. analyzed the performance of onshore wind farms using data mining technologies [14]. Xydas et al. utilized data mining technologies to manage electric vehicle batteries [15]. Apart from the three fields, data mining technologies even had been utilized to combat natural disasters [16]. There are two common types of data mining algorithms, i.e., supervised algorithms and unsupervised algorithms. Compared with the supervised algorithms, the unsupervised algorithms show advantages in the capacity of discovering patterns and association relations among the variables concerned. And, they do not need labeled data which are rare in practice. Therefore, the unsupervised algorithms are more suitable for revealing operational problems of HVAC systems.

Association rule mining has been one of the most promising unsupervised algorithms for analyzing historical operational data of HVAC systems [17]. In the previous studies, many kinds of operational problems (e.g., equipment faults, sensor faults and energy-inefficient operational patterns) had been detected using the association rule mining algorithm in various HVAC systems [18]. Yu et al. introduced the association rule mining to discover energy-inefficient operation patterns and equipment faults of HVAC systems [19]. Then Yu et al. further developed a data mining process for analyzing the operational data of buildings [20]. Xiao et al. also proposed a data mining framework based on the association rule mining to discover abnormal operation patterns hidden in the operational data of HVAC systems [21]. Based on this work, Fan et al. proposed an improved framework for operation performance diagnosis of a central chilling system [22]. They further utilized the method to discover energy conservation opportunities of a HVAC system in a campus building [23,24]. In addition, Fan et al. proposed two temporal association rules mining methods for operation performance diagnosis of HVAC systems [25,26]. Li et al. used the association rule mining to mine the data collected from a variable refrigerant flow system for revealing factors influencing system energy consumption, energy-inefficient control strategies and equipment faults [27]. The association rule mining also had been utilized to analyze the data of other building energy systems and the data of occupant behaviors. Xue et al. adopted the association rule mining to detect sensor faults and energy-inefficient operation patterns in heating systems [28]. Cabrera et al. utilized the association rule mining to identify energy waste patterns of lighting systems [29]. Yu et al. proposed two association rule mining-based methods for identifying energy-inefficient occupant behaviors in residential buildings [30,31]. Rollins et al. proposed a method to discover energy consumption patterns of appliances in residential buildings based on the association rule mining [32]. Wang et al. introduced the association rule mining to analyze impacts of occupant behaviors on residential electricity consumption [33]. D'Oca and Hong identified window opening and closing behaviors of occupants using the association rule mining [34].

However, the existing association rule mining-based methods are still not efficient and not effective enough for discovering operational problems of HVAC systems, which results from two main reasons. Firstly, data preprocessing is always very time-consuming. In general, it costs about 80% of the total time needed for data mining [35]. Outlier identification and data transformation are the most time-consuming steps in the data preprocessing. Outliers are defined as measurements which deviate from the true values significantly. They determine the quality of data transformation [19]. To identify the outliers automatically, some methods had been developed, such as boxplot-based approach [21], Hampel filter [25] and density-based spatial clustering of applications with noise (DBSCAN) [33]. However, the boxplot-based approach is unsuitable to process the data of skewed distribution or thick tailed symmetric distribution [36]. The Hampel filter is not suitable to process the data with obvious data sampling discontinuity. The DBSCAN is difficult to be used since its parameters are usually hard to be determined. It is necessary to develop a proper outlier identification approach for HVAC systems. After outlier identification, data transformation is needed generally. It is because that most association rule mining algorithms such as Apriori [37], Eclat [38] and FP-growth [39] can process categorical data only. But the operational data of HVAC systems are usually numerical. There are two common solutions to transform numerical data into categorical data, i.e., equal-width binning [19] and equal-frequency binning [21]. The equal-width binning divides numerical data into several equal-size intervals. The equal-frequency binning divides numerical data into several categories which contain approximately equal number of measurements. However, these approaches cannot transform data adaptively since they cannot reflect the physical meanings hidden in the data. Secondly, the amount of association rules mined by association rule mining algorithms are always numerous [17]. But, most of the mined association rules are useless for discovering operational problems. For instance, a total of 4272 association rules were obtained by Fan et al., but only 17 valuable association rules were found [22]. It is quite time-consuming to find operational problems from so many association rules. Therefore, it is in great need to develop an automatic approach to discard most of the useless association rules.

In this study, an improved association rule mining-based method is proposed for the first time to discover operational problems of HVAC systems more effectively and more efficiently. A kernel density estimation-based (KDE-based) outlier identification approach is developed to remove outliers automatically. A KDE-based data transformation approach is developed to transform numerical data into categorical data adaptively according to the physical meanings hidden in the data. An association rule comparison-based post mining approach is developed to reduce the amount of useless association rules. Evaluations are made using the historical data of the chiller plant in a commercial building in Shenzhen, China. The computation work is performed using the open-source software Python. The computation tool is a desktop computer with a 3.4 GHz Intel Core i5 processor.

2. Methodology

The outline of the proposed method is as shown in Fig. 1. It consists of three parts, i.e., data preprocessing, association rule mining and association rule comparison-based post mining. The part of data preprocessing aims to improve the quality of the data to be analyzed. It includes four steps, i.e., missing value handling, invalid measurement deleting, KDE-based outlier identification and KDE-based data transformation. Based on the preprocessed data, the purpose of association rule mining is to discover raw association rules which present relations among various variables. The part of association rule comparison-based post mining aims to extract the association rules which are most likely to reveal operational problems from the raw association rules. It includes four steps, i.e., association rule grouping, association rule normalization, association rule comparison and expert analysis.

2.1. Data preprocessing

2.1.1. Missing value handling

The problem of missing values is common due to signal transmission errors, data storage errors, etc. There are many approaches to fill up the missing values such as regression, imputation and inference-based approaches [22]. In this study, only the measurements which have missed for a short time (i.e., less than 1 h) are filled up using linear interpolation approach. If the measurements have missed for a long time (i.e., longer than 1 h), they are discarded because the system operating conditions might have changed significantly.

2.1.2. Invalid measurement deleting

In general, HVAC systems are running during working time and are turned off during nonworking time for non-residential buildings. Measurements collected during the nonworking time are not useful for the detection of operational problems. Therefore, it is better to remove those invalid measurements for the purpose of improving the quality of mined association rules. In this study, the invalid measurements refer to the measurements which are collected during the period when all devices are turned off. The on/off status of a device can be obtained from the on/off measurements directly, or be estimated according to its power, frequency, etc.

2.1.3. KDE-based outlier identification

In the field of statistics, KDE is a popular non-parametric approach to estimate the probability density of a variable [40,41]. It can reveal the distribution characteristics of a variable without any assumptions. In this study, a KDE-based outlier identification approach is developed to identify the outliers. In the previous studies, the outliers are usually defined as the measurements which deviate from the true values significantly. In fact, if the deviations occur frequently, it might indicate that some faults exist in the system or devices. If these outliers are removed, it is impossible to find the reasons behind the outliers. Therefore, in this study, the outliers are identified only when the measurements also have low occurrence frequency.

The schema of the KDE-based outlier identification approach is illustrated in Fig. 2. It includes two steps, i.e., kernel density estimation and outlier identification. In the step of kernel density estimation, the probability density function $f(x)$ of a variable is estimated using KDE algorithm based on its historical measurements as shown in Eq. (1).

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

where x_1, x_2, …, x_n are independent and identically distributed measurements, K is a non-negative kernel function which could be the Gaussian kernel, cosine kernel, triangular kernel, etc., and h is a smoothing parameter named bandwidth.

In the step of outlier identification, a measurement is identified as

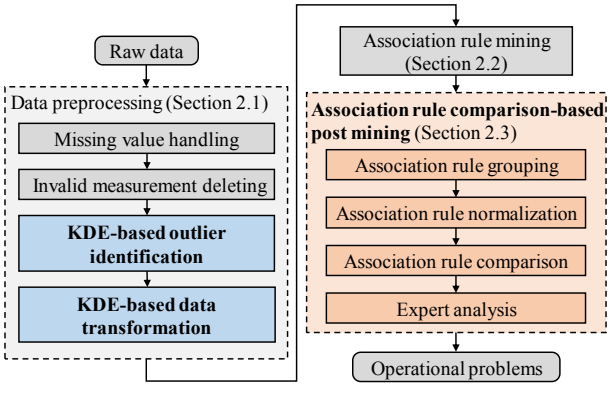


Fig. 1. Outline of the proposed method.

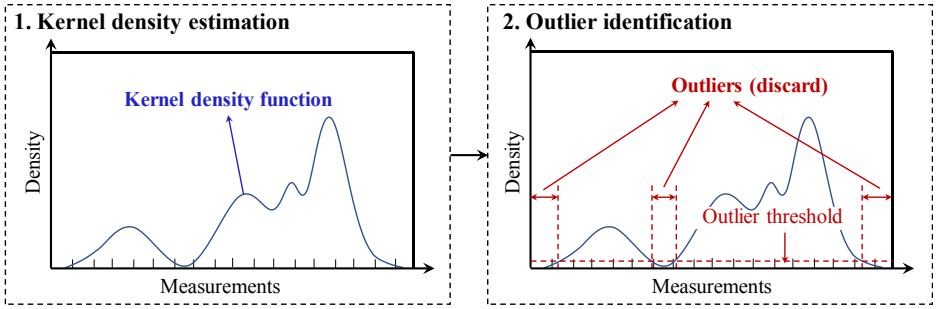


Fig. 2. Schema of the KDE-based outlier identification approach.

an outlier if its estimated density is less than the outlier threshold. The outlier threshold δ_{outlier} is calculated by Eq. (2).

$$\delta_{\text{outlier}} = \frac{f_{\max}}{\alpha} \tag{2}$$

where $f_{\max}$ is the maximum of the estimated probability density function, and α is the factor of the outlier threshold. The outliers are discarded directly in this study.

2.1.4. KDE-based data transformation

A KDE-based data transformation approach is developed to transform numerical data into categorical data. The schema of the KDE-based data transformation approach is illustrated in Fig. 3. It includes three steps, i.e., data classification, category fusion and data transformation. In the step of data classification, based on the results of outlier identification, the remained measurements between two adjacent valleys are grouped into the same category. There might be many categories sometimes because of measurement uncertainty, signal transmission noise and so on. In the step of category fusion, a threshold of the number of categories m is given manually to prevent generating too many categories. If the actual number of categories is more than the threshold of the number of categories, the category with minimum interval is merged with its adjacent category according to its left and right consistency indicators. The left consistency indicator C_{left} and the right consistency indicator C_{right} are defined by Eq. (3) and Eq. (4) respectively. The consistency indicator indicates the similarity between two adjacent categories. Two adjacent categories are similar when the consistency indicator between them is small. Therefore, the minimum-interval category and its right-side category are merged if its C_{right} is smaller than its C_{left}. On the contrary, the minimum-interval category and its left-side category are merged if its C_{right} is bigger than its C_{left}. The step of category fusion is iterated until the number of categories is equal to the threshold of the number of categories. In the last step, measurements in a category are transformed into a unified text form of "X minimum-maximum". The "X" is the variable name of the measurements. The "minimum" is the lower interval boundary of the measurements in the category. And the "maximum" is the upper interval boundary of the measurements in the category.

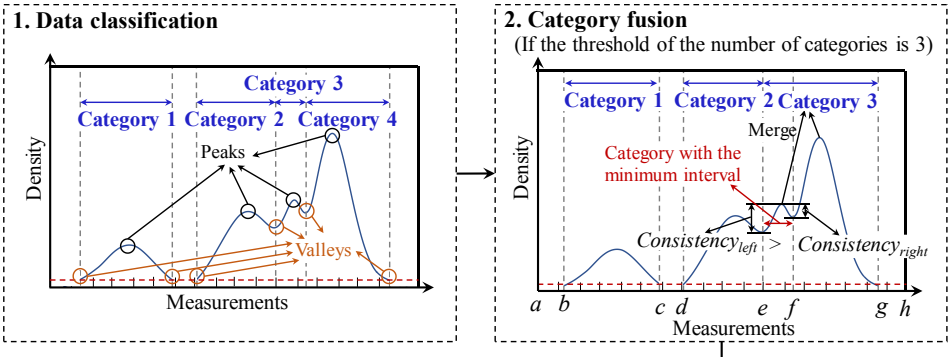
$$C_{\text{left}} = p - v_{\text{left}} \tag{3}$$

$$C_{\text{right}} = p - v_{\text{right}} \tag{4}$$

where p is the peak density of the minimum-interval category, v_{left} is the left-side valley density of the minimum-interval category, and v_{right} is the right-side valley density of the minimum-interval category.

2.2. Association rule mining

Association rule mining is a popular data mining method for discovering association relations among various variables from massive data. An association rule is always in the form of "$A \rightarrow B$" where the A and the B are called the antecedent and the consequent respectively. The performances of different association rule mining algorithms are similar although their association rule mining strategies are various [42]. Therefore, one of the most common association rule mining algorithms named FP-growth is selected in this study [7]. It includes three



3. Data transformation

Category	Transformed measurement*
Category 1	"X *b*-*c*"
Category 2	"X *d*-*e*"
Category 3	"X *e*-*g*"

* "X" is the variable name of the measurements.

Fig. 3. Schema of the KDE-based data transformation approach.

steps. Firstly, the data source is compressed into a special data structure called FP-tree. Secondly, the FP-tree is divided into a set of conditional FP-tree which is a special kind of projected database. Thirdly, each conditional FP-tree is scanned to extract association rules.

The mined association rules are usually filtered preliminarily by three statistical indicators, i.e., support, confidence and lift [7]. The support is an indication of how frequently an association rule appears in a data set. The confidence is an indication of how reliable an association rule is. The lift is an indication of the dependence strength between the antecedent and the consequent of an association rule [22]. Only the association rules whose support, confidence and lift are all greater than the corresponding thresholds are left for further association rule analysis. The support, confidence and lift are defined by Eq. (5), Eq. (6) and Eq. (7) respectively.

$$\text{support}(A \rightarrow B) = P(A \cup B) \tag{5}$$

$$\text{confidence}(A \rightarrow B) = P(B|A) = \frac{P(A \cup B)}{P(A)} \tag{6}$$

$$\text{lift}(A \rightarrow B) = \frac{P(A \cup B)}{P(A)P(B)} \tag{7}$$

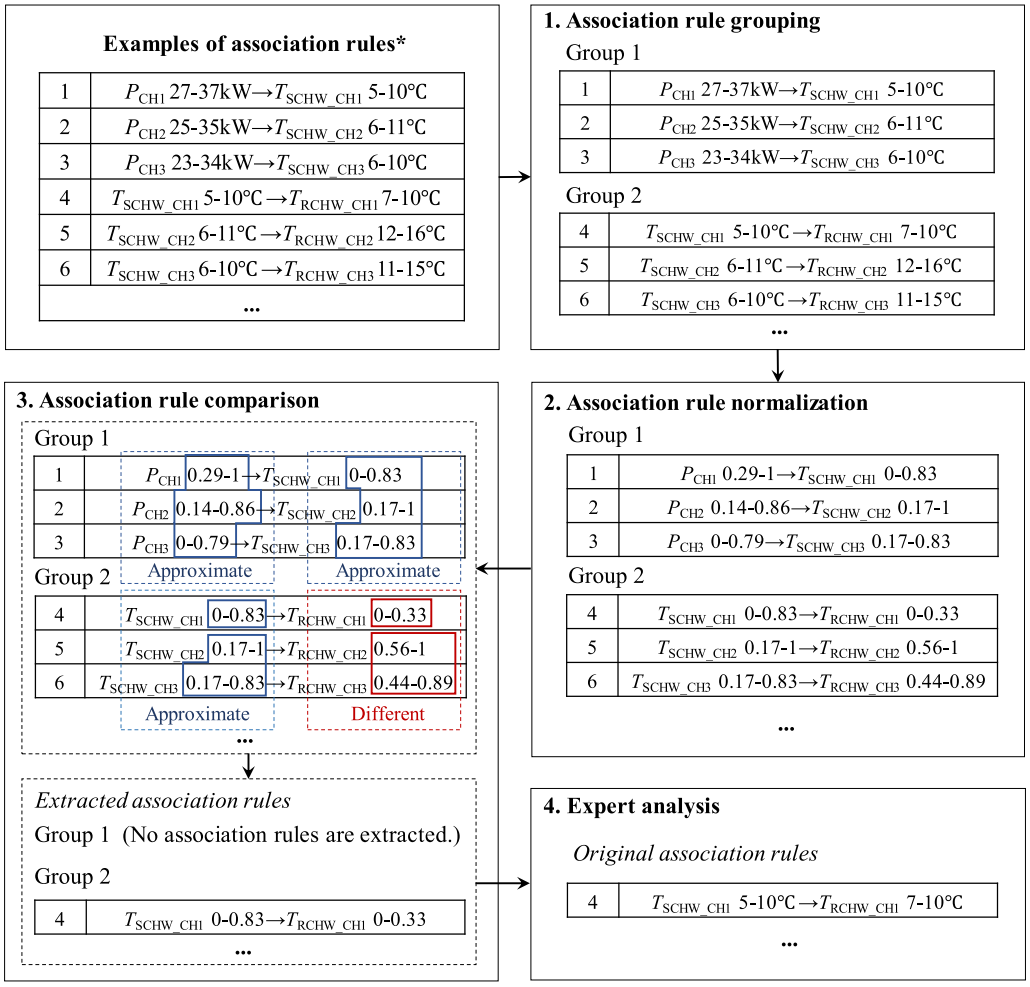
where $P(A \cup B)$ is the probability that the A and the B coincide in the data set to be analyzed, $P(B|A)$ is the conditional probability of the B given the A, $P(A)$ is the probability that the A appears in the data set, and $P(B)$ is the probability that the B appears in the data set.

2.3. Association rule comparison-based post mining

In this study, an association rule comparison-based approach is developed as illustrated in Fig. 4 to extract potentially useful association rules from the numerous raw association rules. The basic idea of this approach is to identify operational problems of a device or a subsystem through comparing its operational patterns with other devices or subsystems which have the same physical meanings as it. It includes four steps, i.e., association rule grouping, association rule normalization, association rule comparison and expert analysis.

2.3.1. Association rule grouping

Association rule grouping aims to classify similar raw association rules into a group. The similar raw association rules refer to the raw association rules that have similar variables in the antecedent, and have similar variables in the consequent. Similar variables are defined as the variables that have the same physical meaning. For instance, the power of the 1# chiller (P_{CH1}) and the power of the 2# chiller (P_{CH2}) have the same physical meaning. The supply chilled water temperature of the 1# chiller (T_{SCHW_CH1}) and the supply chilled water temperature of the 2#



* P_{CHi} is the power of i# chiller

T_{SCHW_CHi} is the supply chilled water temperature of i# chiller

T_{RCHW_CHi} is the return chilled water temperature of i# chiller

Fig. 4. Schema of the association rule comparison-based post mining approach.

chiller (T_{SCHW_CH2}) also have the same physical meaning. Therefore, as the example in Fig. 4, the association rules "P_{CH1} 27–37 kW → T_{SCHW_CH1} 5–10 °C" and "P_{CH2} 25–35 kW → T_{SCHW_CH2} 6–11 °C" are classified into the same group.

2.3.2. Association rule normalization

For the association rules in a group, association rule normalization is utilized to transform the interval boundaries of the similar variables in the association rules into a specific scale. In this study, the min-max normalization is introduced as shown in Eq. (8) [7].

$$z' = \frac{z - z_{min}}{z_{max} - z_{min}} \tag{8}$$

where z is the original lower/upper interval boundary of a variable of an association rule in a group, z' is the normalized lower/upper interval boundary of the variable, z_{max} is the maximum upper interval boundary of all the similar variables in the same group, and z_{min} is the minimum lower interval boundary of all the similar variables in the same group.

2.3.3. Association rule comparison

The step of association rule comparison aims to extract the association rules which are different from other association rules in the same group obviously. In this study, for the similar variables of two association rules in the same group, Euclidean distance is employed to calculate the geometrical distance between their normalized intervals as shown in Eq. (9) [43].

$$D = \sqrt{(l_1 - l_2)^2 + (u_1 - u_2)^2} \tag{9}$$

where, l_1 and u_1 are the lower interval boundary and the upper interval boundary of a variable of one association rule in a group respectively, and l_2 and u_2 are the lower interval boundary and the upper interval boundary of the similar variable of another association rule in the same group respectively.

Based on the distance, two criteria are proposed to define the difference between two association rules in a group. The first one is that the maximum distance between two similar variables in the antecedents of the two association rules exceeds the distance threshold, but the maximum distance between two similar variables in the consequents of the two association rules is less than the distance threshold. The second one is that the maximum distance between two similar variables in the antecedents of the two association rules is less than the distance threshold, but the maximum distance between two similar variables in the consequents of the two association rules exceeds the distance threshold. The distance threshold δ_{dist} is defined by Eq. (10).

$$\delta_{dist} = \beta(n)^{0.5} \tag{10}$$

where β is the factor of distance threshold which is between 0 and 1, and $(n)^{0.5}$ is the theoretical maximum Euclidean distance between two n-dimensional vectors. In this study, the dimension n is 2. In this study, if an association rule is different from 30% of other association rules in the same group, the association rule is extracted as a suspected association rule. The extracted association rules are then analyzed by experts to ultimately discover the operational problems hidden behind them.

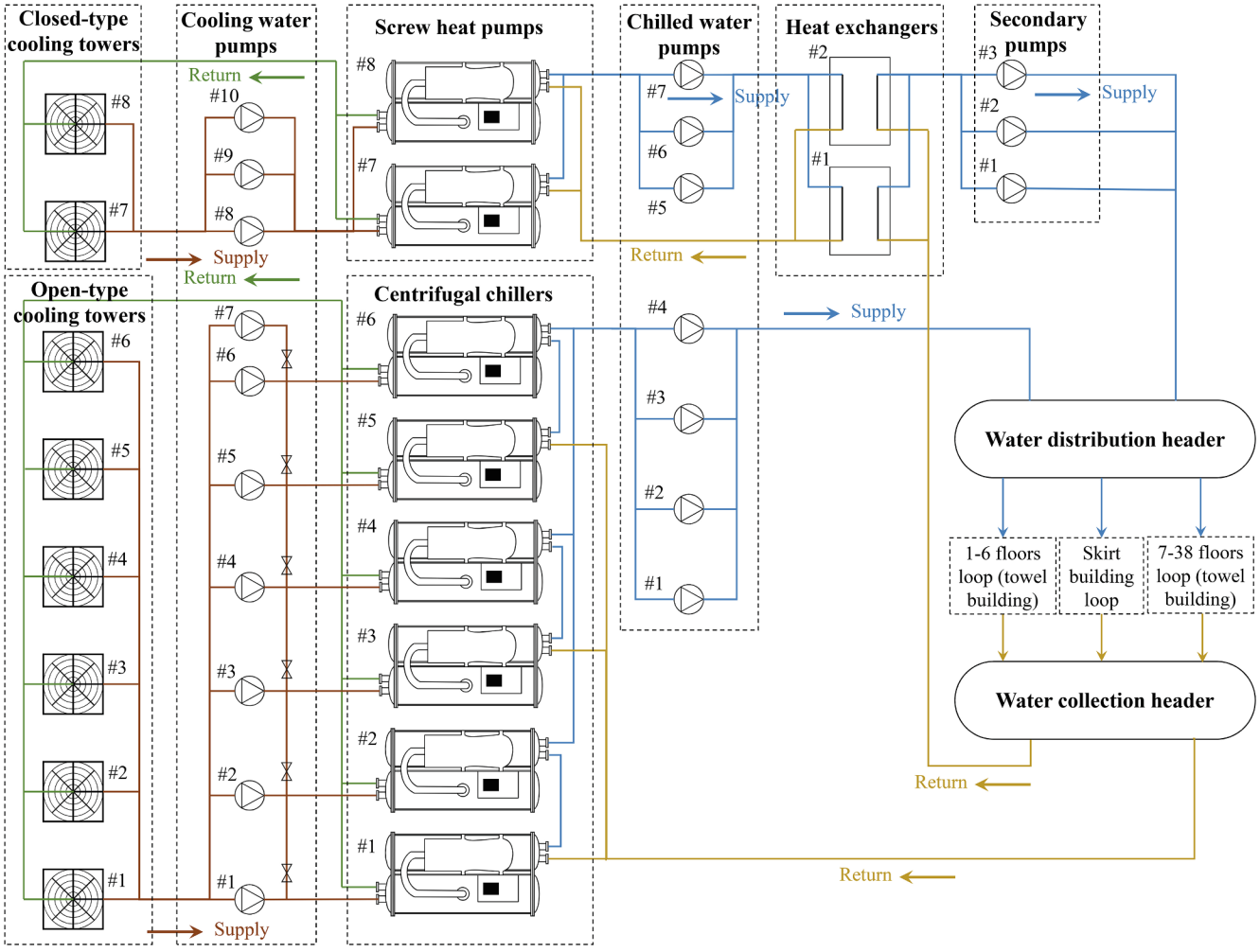


Fig. 5. Schematic diagram of the chiller plant.

3. Evaluations

3.1. Description of the data

Evaluations are made using the data collected from the chiller plant of a commercial building in Shenzhen, China. The commercial building is composed of a skirt building and a tower building. The schematic diagram of the chiller plant is as shown in Fig. 5. The chiller plant contains ten cooling water pumps (COWP), six centrifugal chillers (CH), two screw heat pumps (HP), seven chilled water pumps (CHWP), three secondary pumps (SP), two heat exchangers (HE), a water distribution header (WDH), a water collection header (WCH), and eight cooling towers (CT). There are two types of cooling towers in the chiller plant, i.e., open-type cooling tower (OTCT) and closed-type cooling tower (CTCT). The heat pumps are utilized for cooling in summer and heating in winter. The centrifugal chillers are utilized for cooling only. Two adjacent centrifugal chillers, i.e., 1# and 2# chillers (CH1-2), 3# and 4# chillers (CH3-4), and 5# and 6# chillers (CH5-6), are in series for increasing the chilled water temperature difference in the chiller plant. The water circuit between the screw heat pumps and the heat exchangers is filled with glycol. The water circuit between the closed-type cooling towers and the screw heat pumps is also filled with glycol. The chilled water is distributed into three zones through the water distribution header, i.e., the skirt building (SB), the 1–6 floors of the tower building (TB1-6) and the 7–38 floors of the tower building (TB7-38). A monitoring system is used for collecting the operational data of the chiller plant. The sampling interval is 5 min. One-year historical data collected from May 2016 to May 2017 are selected for the evaluations. The data set includes about 5.4 million observations.

In this study, 49 variables are selected for the purpose of detecting operational problems. All the variables can be classified into 15 types as listed in Table 1. It needs to be noted that, for the 49 variables, the numbers in the name represent the device numbers. For instance, CH1 represent the 1# chiller.

3.2. Results of the data preprocessing

3.2.1. Results of the missing value handling and the invalid measurement deleting

The proportion of the missing values is about 1.27% in the data set. Linear interpolation algorithm is used for filling up the measurements missed for less than 1 h. Invalid measurements are discarded directly which account for about 21.78% of all measurements.

3.2.2. Results of the KDE-based outlier identification and data transformation approaches

In the data set, the measurements of the frequencies and the powers were always 0.0 since the corresponding devices were not running most of the time. Two typical cases are illustrated in Fig. 6(a) and (b). As shown in Fig. 6(a), when the frequency of the 1# cooling water pump is 0.0 Hz, the corresponding peak density is obviously greater than other peak densities. As shown in Fig. 6(b), the power of the 1# chiller also has the same situation as the frequency of the 1# cooling water pump. It makes the KDE-based outlier identification approach unreliable since the factor of outlier threshold is hard to be determined in this situation. Therefore, in this study, for each frequency variable, all measurements with a value of 0.0 are grouped into an individual category directly. It is the same to each power variable. The remained measurements are then handled using the KDE-based outlier identification approach and the KDE-based data transformation approach.

To obtain a reliable density estimation, the bandwidth is optimized in the range between 0.00 and 2.00 for each variable. The factor of outlier threshold α is selected from 50, 100, 150 and 200. For the variables of temperature and relative humidity, the physical meanings of their measurements might be significantly different although the measurements only have a small difference in the values. Therefore, the threshold of the number of categories m is large (i.e., 5) for these variables. m is 3 for the variables of power. It aims to divide the measurements of a variable of power into 3 categories, i.e., “low power”, “medium power” and “high power”. m is not set for the variables of frequency because their set-points are usually discrete in this system. The final parameters utilized in this study are presented in Table 2.

Fig. 7 illustrates the result of the identified outliers and the classified categories of a supply chilled water temperature variable with detail annotations as an instance. Figs. 8 and 9 illustrate the results of the identified outliers and the classified categories of all variables. The annotations of each subgraph in Figs. 8 and 9 refer to Fig. 7. The results show that the developed approach can identify outliers successfully. For instance, for the supply chilled water temperature of the 1# and 2# chillers ($T_{SCHW\_CH1\text{-}2}$), the measurements lower than 5.0 °C are identified as outliers as shown in Fig. 8. For each variable, the outlier proportions are less than 1.10% in this study. The results also show that the classified categories are reasonable. The classified categories of the variables which have the same physical meanings usually look similar. For instance, the supply chilled water temperatures of the chillers, i.e., $T_{SCHW\_CH1\text{-}2}$, $T_{SCHW\_CH3\text{-}4}$, $T_{SCHW\_CH5\text{-}6}$ and T_{TSCHW_CH}, are classified into five similar categories as shown in Fig. 8. Moreover, the categories classified according to the peaks of density distributions can do reflect physical meanings hidden in the data. Most of peaks are usually caused by control logics. For instance, in most of cases, the variables of supply chilled water temperature have a peak between 5.0 °C and 9.0 °C as shown in Fig. 8. It is because that the chilled water temperature set-points are always within this range. Based on the classified categories, the measurements of a variable in a category are transformed into a

Table 1
Selected variables from the database of the chiller plant and their physical meanings.

No.	Type	Variable
1	Outdoor relative humidity	RH_{OUT}
2	Outdoor temperature	T_{OUT}
3	Power of the chiller	P_{CH1}, P_{CH2}, P_{CH3}, P_{CH4}, P_{CH5}, P_{CH6}
4	Power of the heat pump	P_{HP1}, P_{HP2}
5	Frequency of the cooling water pump	F_{COWP1}, F_{COWP2}, F_{COWP3}, F_{COWP4}, F_{COWP5}, F_{COWP6}, F_{COWP7}
6	Frequency of the chilled water pump	F_{CHWP1}, F_{CHWP2}, F_{CHWP3}, F_{CHWP4}
7	Frequency of the open-type cooling tower fan	F_{CT1}, F_{CT2}, F_{CT3}, F_{CT4}, F_{CT5}, F_{CT6}
8	Supply chilled water temperature	$T_{SCHW\_CH1\text{-}2}$, $T_{SCHW\_CH3\text{-}4}$, $T_{SCHW\_CH5\text{-}6}$, T_{TSCHW_CH}, T_{TSCHW_HE}, T_{TSCHW_WDH}
9	Return chilled water temperature	$T_{RCHW\_CH1\text{-}2}$, $T_{RCHW\_CH3\text{-}4}$, $T_{RCHW\_CH5\text{-}6}$, T_{TRCHW_CH}, T_{TRCHW_HE}, T_{RCHW_SB}, $T_{RCHW\_TB1\text{-}6}$, $T_{RCHW\_TB7\text{-}38}$
10	Supply glycol temperature of the heat pump	T_{SG_HP1}, T_{SG_HP2}
11	Return glycol temperature of the heat pump	T_{RG_HP1}, T_{RG_HP2}
12	Total supply glycol temperature of the closed-type cooling towers	T_{TSG_CTCT}
13	Total return glycol temperature of the closed-type cooling towers	T_{TRG_CTCT}
14	Total supply cooling water temperature of the open-type cooling towers	T_{TSCOW_OTCT}
15	Total return cooling water temperature of the open-type cooling towers	T_{TRCOW_OTCT}

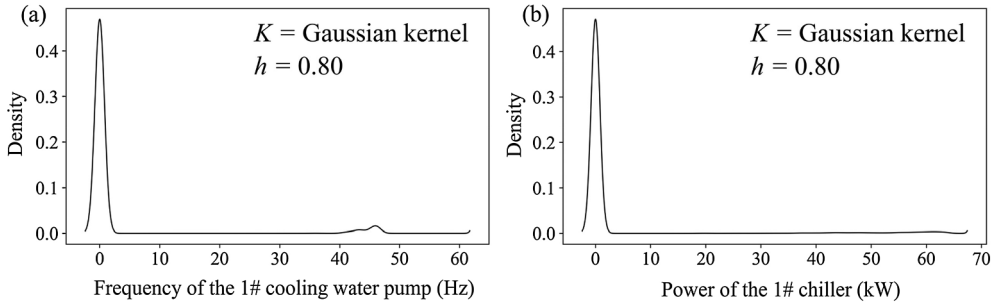


Fig. 6. Density distributions of two typical variables.

Table 2
Parameters of the KDE-based data transformation approach.

Variable type	Kernel function K	Bandwidth h	Factor of outlier threshold α	Threshold of the number of categories m
Temperature	Gaussian kernel	Given in Figs. 8 and 9	100	5
Relative humidity				5
Frequency				–
Power				3

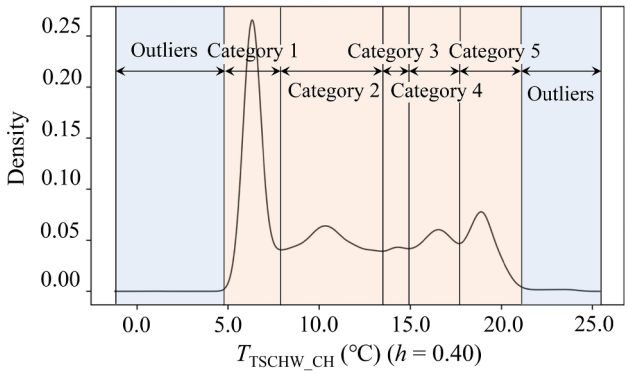


Fig. 7. The identified outliers and the classified categories of a supply chilled water temperature variable (h is the bandwidth).

specific form which can indicate the variable name and the interval boundaries of the category, e.g., "RH_{OUT} 42.3–69.1%", "T_{OUT} 8.3–14.3 °C".

3.3. Results of the association rule comparison-based post mining approach

In this study, the threshold of the support is set to 1.00% to remove the very infrequent association rules only. The threshold is low because some association rules have rather low support but meaningful for revealing operational problems. For instance, in this study, the support of the association rule "F_{COWP5} 33.1–37.8 Hz → F_{COWP6} 38.3–44.1 Hz" is only 1.06%. But it indicates an abnormal operation pattern of the cooling water pumps. The thresholds of the confidence and the lift are set to 40.00% and 1.00 respectively to remove the association rules which are not strong. Since the amount of association rules are numerous, this study analyzes the association rules which have one variable in the antecedent and one variable in the consequent respectively only. There are a total of 5800 raw association rules to be further analyzed in this study.

All the association rules are grouped according to the physical meanings of the variables in the antecedent and the consequent. The physical meanings of the variables are listed in Table 1. Association rules in a group should have variables of the same kind of physical meanings in the antecedents, and have variables of the same kind of physical meanings in the consequents. Based on this principle, all the association rules are classified into 217 groups.

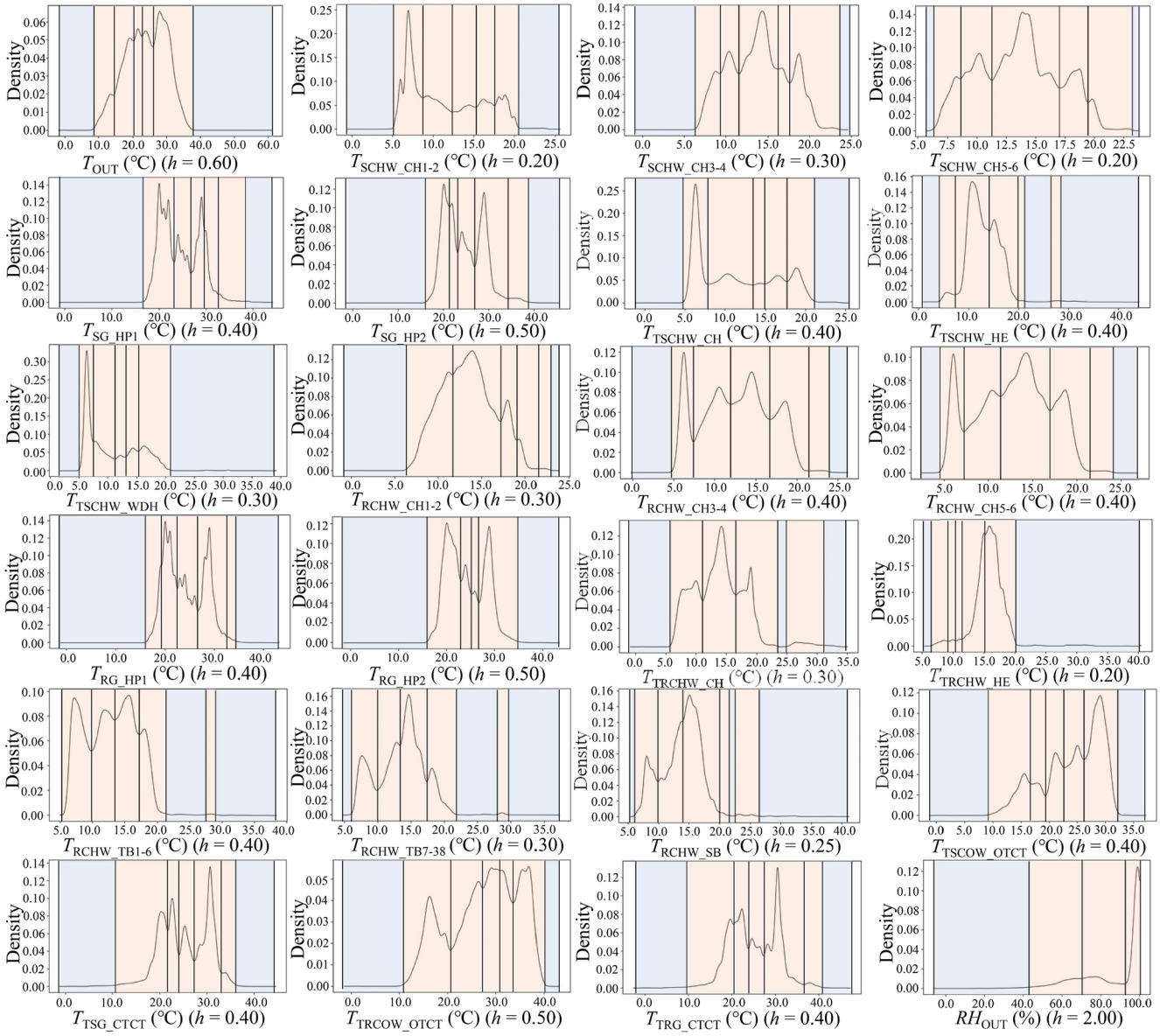
Each association rule is compared with all the other association rules in the same group. The factor of distance threshold β is 0.4 in this study. Finally, 54.98% of the raw association rules are discarded. A total of 2611 suspected association rules are left for further analysis by experts. After checking all the suspected association rules carefully, a total of 84 abnormal association rules are discovered finally. Among them, three discovered operational problems of different kinds are discussed as instances.

3.3.1. Case 1: Faults of supply and return chilled water temperature sensors

Four abnormal association rules related to the faults of supply chilled water temperature sensors are found from the suspected association rules, as listed in Table 3. To reveal the differences among the abnormal association rules and the normal association rules, eight normal association rules are listed in Table 4. The normal association rules reveal that the supply chilled water temperatures and the total supply chilled water temperature were usually lower than 9.0 °C when the corresponding chillers were running. However, the abnormal association rules in the Table 3 indicate that the supply chilled water temperatures of the 3-4# chillers and the 5-6# chillers (i.e., T_{SCHW_CH3-4} and T_{SCHW_CH5-6}) were usually higher than 11.0 °C when the corresponding chillers were running. It can be concluded that some faults occurred in the supply chilled water temperature sensors of the 3-4# chillers and the 5-6# chillers.

Four abnormal association rules related to the faults of return chilled water temperature sensors are found, as listed in Table 5. To reveal the differences among the abnormal association rules and the normal association rules, eight normal association rules are listed in Table 6 for comparisons. The normal association rules reveal that the return chilled water temperatures and the total return chilled water temperature were usually higher than 11.0 °C when the corresponding chillers were running. However, the abnormal association rules in the Table 5 indicate that the return chilled water temperatures of the 3-4# chillers and the 5-6# chillers (i.e., T_{RCHW_CH3-4} and T_{RCHW_CH5-6}) were usually lower than 7.5 °C when the corresponding chillers were running. It can be deduced that some faults occurred in the return chilled water temperature sensors of the 3-4# chillers and the 5-6# chillers.

Fig. 10(a) illustrates the supply and return chilled water temperatures of the 3-4# chillers (i.e., T_{SCHW_CH3-4} and T_{RCHW_CH3-4}) as well as the total supply and return chilled water temperatures of chillers (i.e., T_{TSCHW_CH} and T_{TRCHW_CH}) on August 11, 2016. It shows that, for the 3-4# chillers, the supply chilled water temperature of them was very similar to the total return chilled water temperature of chillers. And the return chilled water temperature of them was very similar to the total supply chilled water temperature of chillers. Fig. 10(b) illustrates the supply and return chilled water temperatures of the 5-6# chillers (i.e., T_{SCHW_CH5-6} and T_{RCHW_CH5-6}) as well as the total supply and return chilled water temperatures of chillers (i.e., T_{TSCHW_CH} and T_{TRCHW_CH}) on August 9, 2016. It shows the same fault as the Fig. 10(a). After consulting with the technicians, they found that the supply chilled



Note: Annotations of each subgraph refer to Fig. 7.

Fig. 8. The identified outliers and the classified categories of all the temperature variables and the relative humidity variable (h is the bandwidth).

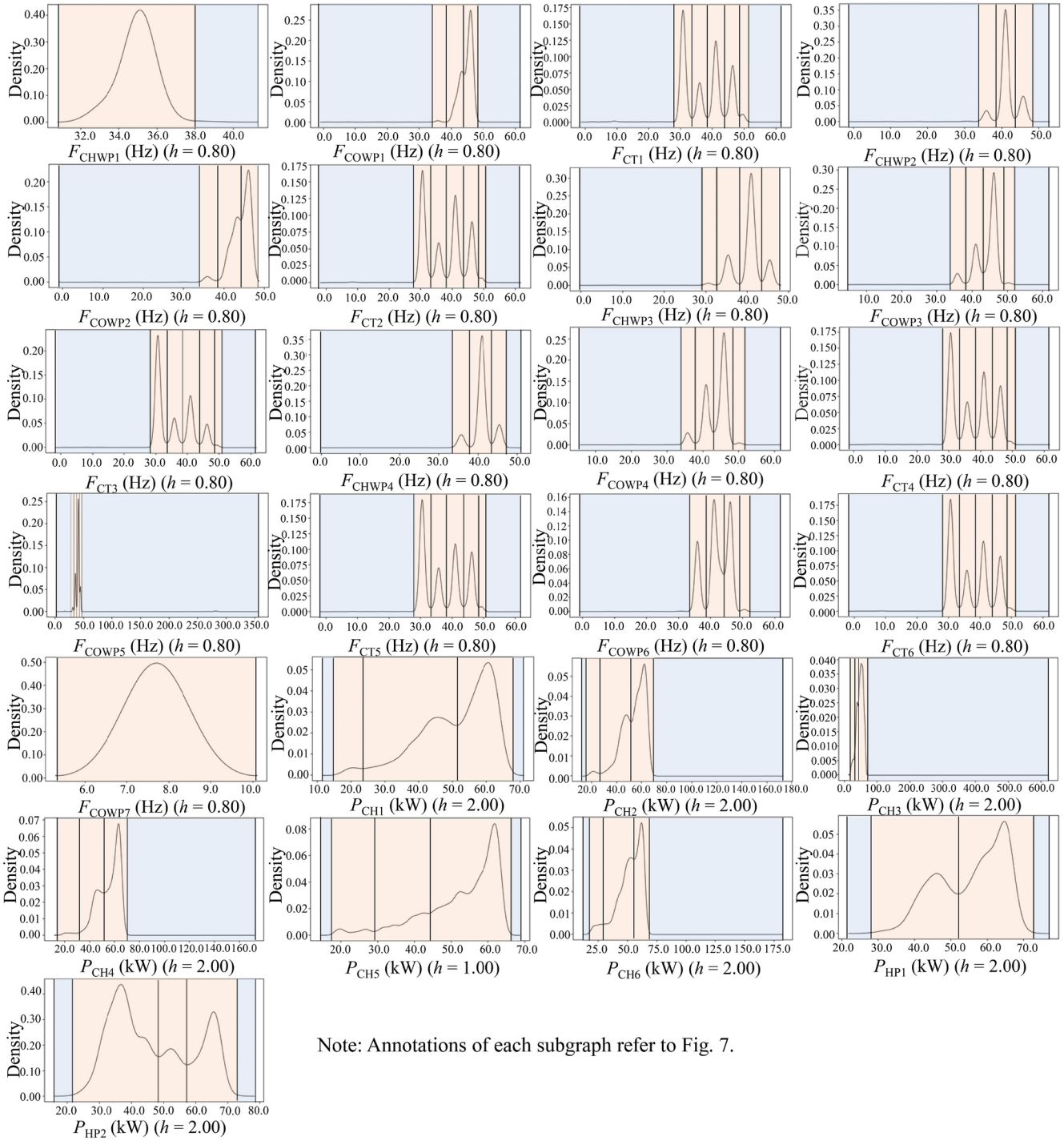
water temperature sensors were labeled as the return chilled water temperature sensors for the 3-4# chillers and the 5-6# chillers in the monitoring system. And the return chilled water temperature sensors were labeled as the supply chilled water temperature sensors for the 3-4# chillers and the 5-6# chillers in the monitoring system.

3.3.2. Case 2: Abnormal operation patterns related to cooling water pumps

Association rules can also reveal operation patterns of HVAC systems. Taking the operation of pumps for example, if pumps are in parallel, they should have the same frequency. Table 8 provides eleven normal association rules as instances. They indicate that the frequencies of two cooling water pumps were usually the same. Three abnormal association rules are found from the suspected association rules, as listed in Table 7. They reveal that the frequency of the 5# cooling water pump (i.e., F_{COWP5}) was often different with the frequency of the 6# cooling water pump (i.e., F_{COWP6}). After checking the historical data carefully, it was found that the frequency of the 5# cooling water pump was always about 5.0 Hz smaller than the frequency of the 6# cooling water pump until September 1, 2016. The frequency of the two pumps were similar after September 1, 2016. After consulting with the technicians, the control strategy of frequencies of the two pumps was wrong before September 1, 2016. And it was discovered and modified after September 1, 2016.

3.3.3. Case 3: Abnormal operation patterns related to chilled water distributions

Two abnormal association rules are found which reveal the abnormal operation patterns of return chilled water, as listed in Table 9. Eleven normal association rules are also provided for comparisons as listed in Table 10. In general, in an energy efficient chilled water system, the return chilled water temperatures of different chilled water circuits should be almost the same. However, the abnormal association rules indicate that the return chilled water temperature of the 1–6 floors of the tower building (i.e., $T_{RCHW\_TB1\text{-}6}$) was significantly lower than the return chilled water temperatures of the skirt building (i.e., T_{RCHW_SB}) and the 7–38 floors of the tower building (i.e., $T_{RCHW\_TB7\text{-}38}$) sometimes.



Note: Annotations of each subgraph refer to Fig. 7.

Fig. 9. The identified outliers and the classified categories of all the frequency variables and all the power variables (h is the bandwidth).

Fig. 11 illustrates the abnormal patterns on May 26, 2016. $T_{RCHW\_TB1\text{-}6}$ was obviously lower than T_{RCHW_SB} and $T_{RCHW\_TB7\text{-}38}$ from 8:45 to 11:10 and from 13:50 to 17:40. However, $T_{RCHW\_TB1\text{-}6}$ was close to T_{RCHW_SB} and $T_{RCHW\_TB7\text{-}38}$ from 7:20 to 8:45 and from 11:10 to 13:40. It also increased from 17:40 to 19:00. The authors found that there were mainly dining rooms and fitness rooms on the 1–6 floors of

Table 3
Abnormal association rules related to the faults of supply chilled water temperature sensors.

No.	Antecedent	Consequent	Support	Confidence	Lift
1	P_{CH3} 41.6–69.4 kW	$T_{SCHW\_CH3\text{-}4}$ 11.5–16.2 °C	8.57%	92.10%	2.03
2	P_{CH4} 50.9–70.6 kW	$T_{SCHW\_CH3\text{-}4}$ 11.5–16.2 °C	8.88%	90.32%	1.99
3	P_{CH5} 43.7–66.4 kW	$T_{SCHW\_CH5\text{-}6}$ 11.3–17.1 °C	6.58%	94.01%	1.76
4	P_{CH6} 55.7–69.6 kW	$T_{SCHW\_CH5\text{-}6}$ 11.3–17.1 °C	5.34%	85.66%	1.61

Table 4
Normal association rules related to the supply chilled water temperature sensors.

No.	Antecedent	Consequent	Support	Confidence	Lift
1	P_{CH1} 51.1–68.9 kW	$T_{SCHW\_CH1\text{-}2}$ 5.0–8.7 °C	2.90%	97.45%	2.48
2	P_{CH2} 51.5–69.9 kW	$T_{SCHW\_CH1\text{-}2}$ 5.0–8.7 °C	4.61%	90.42%	2.31
3	P_{CH1} 51.1–68.9 kW	T_{TSCHW_CH} 4.8–7.9 °C	2.80%	94.05%	2.63
4	P_{CH2} 51.5–69.9 kW	T_{TSCHW_CH} 4.8–7.9 °C	4.40%	86.29%	2.42
5	P_{CH3} 41.6–69.4 kW	T_{TSCHW_CH} 4.8–7.9 °C	8.89%	95.52%	2.67
6	P_{CH4} 50.9–70.6 kW	T_{TSCHW_CH} 4.8–7.9 °C	9.09%	92.39%	2.59
7	P_{CH5} 43.7–66.4 kW	T_{TSCHW_CH} 4.8–7.9 °C	6.65%	95.09%	2.66
8	P_{CH6} 55.7–69.6 kW	T_{TSCHW_CH} 4.8–7.9 °C	5.09%	81.66%	2.29

Table 5
Abnormal association rules related to the faults of return chilled water temperature sensors.

No.	Antecedent	Consequent	Support	Confidence	Lift
1	P_{CH3} 41.6–69.4 kW	$T_{RCHW\_CH3\text{-}4}$ 4.7–7.4 °C	8.82%	94.73%	6.37
2	P_{CH4} 50.9–70.6 kW	$T_{RCHW\_CH3\text{-}4}$ 4.7–7.4 °C	8.98%	91.33%	6.14
3	P_{CH5} 43.7–66.4 kW	$T_{RCHW\_CH5\text{-}6}$ 4.5–7.2 °C	6.54%	93.49%	6.67
4	P_{CH6} 55.7–69.6 kW	$T_{RCHW\_CH5\text{-}6}$ 4.5–7.2 °C	4.87%	78.15%	5.58

Table 6
Normal association rules related to the return chilled water temperature sensors.

No.	Antecedent	Consequent	Support	Confidence	Lift
1	P_{CH1} 51.1–68.9 kW	$T_{RCHW\_CH1\text{-}2}$ 11.6–17.1 °C	2.89%	97.01%	1.74
2	P_{CH2} 51.5–69.9 kW	$T_{RCHW\_CH1\text{-}2}$ 11.6–17.1 °C	3.89%	76.21%	1.37
3	P_{CH1} 51.1–68.9 kW	T_{TRCHW_CH} 11.1–16.6 °C	2.83%	95.06%	1.97
4	P_{CH2} 51.5–69.9 kW	T_{TRCHW_CH} 11.1–16.6 °C	3.94%	77.25%	1.60
5	P_{CH3} 41.6–69.4 kW	T_{TRCHW_CH} 11.1–16.6 °C	8.67%	93.14%	1.93
6	P_{CH4} 50.9–70.6 kW	T_{TRCHW_CH} 11.1–16.6 °C	9.10%	92.58%	1.91
7	P_{CH5} 43.7–66.4 kW	T_{TRCHW_CH} 11.1–16.6 °C	6.50%	92.87%	1.92
8	P_{CH6} 55.7–69.6 kW	T_{TRCHW_CH} 11.1–16.6 °C	5.21%	83.67%	1.73

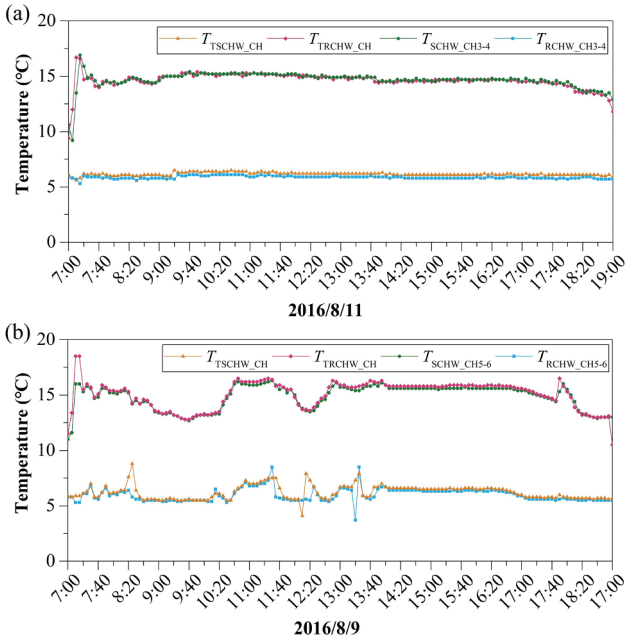


Fig. 10. Supply and return chilled water temperatures of (a) the 3-4# chillers and (b) the 5-6# chillers as well as the total supply and return chilled water temperatures of chillers on two typical days.

Table 7
Abnormal association rules related to the abnormal operation patterns of cooling water pumps.

No.	Antecedent	Consequent	Support	Confidence	Lift
1	F_{COWP5} 33.1–37.8 Hz	F_{COWP6} 38.3–44.1 Hz	1.06%	58.89%	9.05
2	F_{COWP6} 44.1–48.6 Hz	F_{COWP5} 37.9–43.3 Hz	2.80%	59.72%	9.89
3	F_{COWP5} 37.9–43.3 Hz	F_{COWP6} 44.1–48.6 Hz	2.80%	46.34%	9.89

Table 8
Normal association rules related to the normal operation patterns of cooling water pumps.

No.	Antecedent	Consequent	Support	Confidence	Lift
1	F_{COWP1} 39.1–44.0 Hz	F_{COWP2} 38.1–44.1 Hz	2.24%	94.80%	24.25
2	F_{COWP2} 38.1–44.1 Hz	F_{COWP1} 39.1–44.0 Hz	2.24%	57.31%	24.25
3	F_{COWP4} 37.9–43.0 Hz	F_{COWP3} 38.1–43.1 Hz	3.19%	59.96%	17.56
4	F_{COWP3} 38.1–43.1 Hz	F_{COWP4} 37.9–43.0 Hz	3.19%	93.40%	17.56
5	F_{COWP6} 38.3–44.1 Hz	F_{COWP5} 37.9–43.3 Hz	2.74%	42.15%	6.98
6	F_{COWP5} 37.9–43.3 Hz	F_{COWP6} 38.3–44.1 Hz	2.74%	45.40%	6.98
7	F_{COWP5} 43.3–46.5 Hz	F_{COWP6} 44.1–48.6 Hz	1.03%	94.87%	20.24
8	F_{COWP2} 44.1–48.6 Hz	F_{COWP1} 44.0–48.5 Hz	3.43%	75.18%	21.72
9	F_{COWP1} 44.0–48.5 Hz	F_{COWP2} 44.1–48.6 Hz	3.43%	99.09%	21.72
10	F_{COWP3} 43.1–48.8 Hz	F_{COWP4} 43.0–48.5 Hz	8.05%	89.65%	10.22
11	F_{COWP4} 43.0–48.5 Hz	F_{COWP3} 43.1–48.8 Hz	8.05%	91.78%	10.22

Table 9
Abnormal association rules related to the abnormal chilled water distributions.

No.	Antecedent	Consequent	Support	Confidence	Lift
1	$T_{RCHW\_TB1\text{-}6}$ 9.9–13.4 °C	$T_{RCHW\_TB7\text{-}38}$ 13.3–17.4 °C	13.27%	49.51%	1.15
2	$T_{RCHW\_TB1\text{-}6}$ 9.9–13.4 °C	T_{RCHW_SB} 13.9–20.3 °C	14.60%	54.49%	1.03

the tower building. The occupant densities of these floors significantly decreased during non-mealtimes, which caused the decreases of the cooling load. But the flow rate of the supply chilled water was not adjusted according to the changes of cooling load in this system.

4. Conclusions

Nowadays, massive amounts of operational data of HVAC systems are collected in buildings. It is very valuable to reveal operational problems from these data for improving the operational performance of HVAC systems. Association rule mining is a promising data mining technology to achieve this goal. However, it is quite time-consuming to preprocess data and extract operational problems from numerous association rules in the previous studies. This paper proposes an improved association rule mining-based method to discover the operational problems of HVAC systems more effectively and more efficiently. A kernel density estimation-based outlier identification approach is developed to remove outliers automatically. A kernel density estimation-based data transformation approach is developed to transform numerical data into categorical data adaptively according to the physical meanings hidden in the data. An association rule comparison-based post mining approach is developed to extract the suspected association rules which are most likely to reveal operational problems from the raw association rules automatically. It is helpful to reduce useless association rules effectively.

Evaluations are made using the one-year historical data of 49 variables collected from the chiller plant of a commercial building in Shenzhen, China. Results show that the kernel density estimation-based outlier identification approach can detect the outliers of the measurements of the 49 variables effectively. In this case, the outlier proportions of all the variables are less than 1.10%. Results further show that

Table 10
Normal association rules related to the normal chilled water distributions.

No.	Antecedent	Consequent	Support	Confidence	Lift
3	$T_{RCHW\_TB7\text{-}38}$ 17.4–21.5 °C	T_{RCHW_SB} 13.9–20.3 °C	11.34%	87.91%	1.66
4	$T_{RCHW\_TB1\text{-}6}$ 17.2–21.6 °C	T_{RCHW_SB} 13.9–20.3 °C	12.82%	89.49%	1.69
5	$T_{RCHW\_TB7\text{-}38}$ 17.4–21.5 °C	$T_{RCHW\_TB1\text{-}6}$ 17.2–21.6 °C	10.65%	82.53%	5.76
6	$T_{RCHW\_TB1\text{-}6}$ 17.2–21.6 °C	$T_{RCHW\_TB7\text{-}38}$ 17.4–21.5 °C	10.65%	74.35%	5.76
7	$T_{RCHW\_TB7\text{-}38}$ 17.4–21.5 °C	T_{TRCHW_CH} 16.6–23.2 °C	12.55%	97.29%	4.35
8	T_{TRCHW_CH} 16.6–23.2 °C	$T_{RCHW\_TB7\text{-}38}$ 17.4–21.5 °C	12.55%	56.09%	4.35
9	$T_{RCHW\_TB1\text{-}6}$ 17.2–21.6 °C	T_{TRCHW_CH} 16.6–23.2 °C	11.73%	81.88%	3.66
10	T_{TRCHW_CH} 16.6–23.2 °C	$T_{RCHW\_TB1\text{-}6}$ 17.2–21.6 °C	11.73%	52.41%	3.66
11	T_{TRCHW_CH} 16.6–23.2 °C	T_{RCHW_SB} 13.9–20.3 °C	19.86%	88.75%	1.67

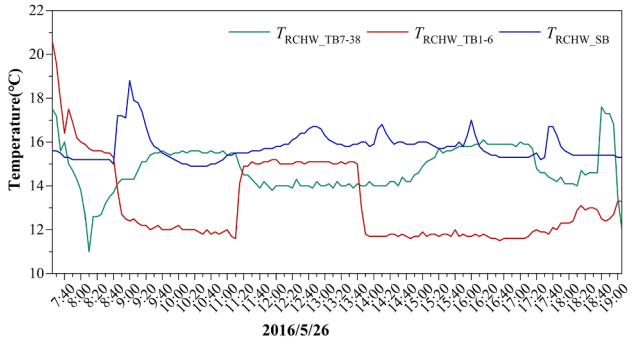


Fig. 11. Return chilled water temperatures of the 7–38 floors of the tower building, the 1–6 floors of the tower building and the skirt building on a typical day.

the kernel density estimation-based data transformation approach can categorize the numerical measurements of every variable properly. The developed association rule comparison-based post mining approach can filter out a large number of useless association rules which account for 54.98% of the raw association rules. After the post mining, there are 2611 association rules left. In the end, 84 abnormal association rules are discovered from the remained association rules after analyzing every association rule carefully. They reveal operational problems of various kinds, e.g., faults of temperature sensors, abnormal operation of pumps and energy-inefficient chilled water distributions.

The proposed method provides a generic solution to detect operational problems of HVAC systems from their historical operational data effectively. It can be applied to improve the operational performance of buildings which have numerous historical operational data. The proposed method might be also useful to detect operational problems from the historical operational data of district heating systems, district cooling systems, distributed energy systems and so on. Future studies are suggested to enhance the capacity in filtering out useless association rules.

Acknowledgements

This study is supported by the National Natural Science Foundation of China (Grant No. 51706197).

References

[1] Fan C, Xiao F, Zhao Y. A short-term building cooling load prediction method using deep learning algorithms. Appl Energy 2017;195:222–33. https://doi.org/10.1016/j.apenergy.2017.03.064.
[2] Hong T, Yang L, Hill D, Feng W. Data and analytics to inform energy retrofit of high performance buildings. Appl Energy 2014;126:90–106. https://doi.org/10.1016/j.apenergy.2014.03.052.
[3] Tong Z, Chen Y, Malkawi A. Estimating natural ventilation potential for high-rise buildings considering boundary layer meteorology. Appl Energy 2017;193:276–86. https://doi.org/10.1016/j.apenergy.2017.02.041.
[4] Tong Z, Chen Y, Malkawi A, Liu Z, Freeman RB. Energy saving potential of natural ventilation in China: the impact of ambient air pollution. Appl Energy 2016;179:660–8. https://doi.org/10.1016/j.apenergy.2016.07.019.
[5] Chen Y, Tong Z, Wu W, Samuelson H, Malkawi A, Norford L. Achieving natural ventilation potential in practice: control schemes and levels of automation. Appl Energy 2019;235:1141–52. https://doi.org/10.1016/j.apenergy.2018.11.016.
[6] Katipamula S, Brambley MR. Review article: methods for fault detection, diagnostics, and prognostics for building systems—a review, part I. HVAC&R Res 2005;11:3–25. https://doi.org/10.1080/10789669.2005.10391123.
[7] Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Waltham, USA: Morgan Kaufmann; 2012https://doi.org/10.1016/C2009-0-61819-5.
[8] Pérez-Martín A, Pérez-Torregrosa A, Vaca M. Big data techniques to measure credit banking risk in home equity loans. J Bus Res 2018;89:448–54. https://doi.org/10.1016/j.jbusres.2018.02.008.
[9] Chen MC, Chiu AL, Chang HH. Mining changes in customer behavior in retail marketing. Expert Syst Appl 2005;28:773–81. https://doi.org/10.1016/j.eswa.2004.12.033.
[10] Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technol Forecast Soc Chang 2018;126:3–13. https://doi.org/10.1016/j.techfore.2015.12.019.
[11] Torregrossa D, Hansen J, Hernández-Sancho F, Cornelissen A, Schutz G, Leopold U. A data-driven methodology to support pump performance analysis and energy efficiency optimization in Waste Water Treatment Plants. Appl Energy 2017;208:1430–40. https://doi.org/10.1016/j.apenergy.2017.09.012.
[12] Zhang Z, Kusiak A, Zeng Y, Wei X. Modeling and optimization of a wastewater pumping system with data-mining methods. Appl Energy 2016;164:303–11. https://doi.org/10.1016/j.apenergy.2015.11.061.
[13] Heng J, Wang J, Xiao L, Lu H. Research and application of a combined model based on frequent pattern growth algorithm and multi-objective optimization for solar radiation forecasting. Appl Energy 2017;208:845–66. https://doi.org/10.1016/j.apenergy.2017.09.063.
[14] Astolfi D, Castellani F, Garinei A, Terzi L. Data mining techniques for performance analysis of onshore wind farms. Appl Energy 2015;148:220–33. https://doi.org/10.1016/j.apenergy.2015.03.075.
[15] Xydas E, Marmaras C, Cipcigan LM, Jenkins N, Carroll S, Barker M. A data-driven approach for characterising the charging demand of electric vehicles: a UK case study. Appl Energy 2016;162:763–71. https://doi.org/10.1016/j.apenergy.2015.10.151.
[16] Goswami S, Chakraborty S, Ghosh S, Chakrabarti A, Chakraborty B. A review on application of data mining techniques to combat natural disasters. Ain Shams Eng J 2018;9:365–78. https://doi.org/10.1016/j.asej.2016.01.012.
[17] Fan C, Xiao F, Li Z, Wang J. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: a review. Energy Build 2018;159:296–308. https://doi.org/10.1016/j.enbuild.2017.11.008.
[18] Zhao Y, Li T, Zhang X, Zhang C. Artificial intelligence-based fault detection and diagnosis methods for building energy systems: advantages, challenges and the future. Renew Sustain Energy Rev 2019;109:85–101. https://doi.org/10.1016/j.rser.2019.04.021.
[19] Yu Z, Haghighat F, Fung BCM, Zhou L. A novel methodology for knowledge discovery through mining associations between building operational data. Energy Build 2012;47:430–40. https://doi.org/10.1016/j.enbuild.2011.12.018.
[20] Yu Z, Fung BCM, Haghighat F. Extracting knowledge from building-related data—a data mining framework. Build Simul 2013;6:207–22. https://doi.org/10.1007/s12273-013-0117-8.
[21] Xiao F, Fan C. Data mining in building automation system for improving building operational performance. Energy Build 2014;75:109–18. https://doi.org/10.1016/j.enbuild.2014.02.005.
[22] Fan C, Xiao F, Yan C. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Autom Constr 2015;50:81–90. https://doi.org/10.1016/j.autcon.2014.12.006.
[23] Fan C, Xiao F. Assessment of building operational performance using data mining techniques: a case study. Energy Proc 2017;111:1070–8. https://doi.org/10.1016/j.egypro.2017.03.270.
[24] Fan C, Xiao F. Mining big building operational data for improving building energy efficiency: a case study. Build Serv Eng Res Technol 2018;39:117–28. https://doi.org/10.1177/0143624417704977.
[25] Fan C, Xiao F, Madsen H, Wang D. Temporal knowledge discovery in big BAS data for building energy management. Energy Build 2015;109:75–89. https://doi.org/10.1016/j.enbuild.2015.09.060.
[26] Fan C, Sun Y, Shan K, Xiao F, Wang J. Discovering gradual patterns in building operations for improving building energy efficiency. Appl Energy

2018;224:116–23. https://doi.org/10.1016/j.apenergy.2018.04.118.
[27] Li G, Hu Y, Chen H, Li H, Hu M, Guo Y, et al. Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions. Appl Energy 2017;185:846–61. https://doi.org/10.1016/j.apenergy.2016.10.091.
[28] Xue P, Zhou Z, Fang X, Chen X, Liu L, Liu Y, et al. Fault detection and operation optimization in district heating substations based on data mining techniques. Appl Energy 2017;205:926–40. https://doi.org/10.1016/j.apenergy.2017.08.035.
[29] Cabrera DFM, Zareipour H. Data association mining for identifying lighting energy waste patterns in educational institutes. Energy Build 2013;62:210–6. https://doi.org/10.1016/j.enbuild.2013.02.049.
[30] Yu Z, Haghighat F, Fung BCM, Yoshino H, Morofsky E. A methodology for identifying and improving occupant behavior in residential buildings. Energy 2011;36:6596–608. https://doi.org/10.1016/j.energy.2011.09.002.
[31] Yu Z, Li J, Li HQ, Han J, Zhang GQ. A novel methodology for identifying associations and correlations between household appliance behaviour in residential buildings. Energy Proc 2015;78:591–6. https://doi.org/10.1016/j.egypro.2015.11.024.
[32] Rollins S, Banerjee N. Using rule mining to understand appliance energy consumption patterns. In: 2014 IEEE international conference on pervasive computing and communications; 2014. p. 29–37. https://doi.org/10.1109/PerCom.2014.6813940.
[33] Wang F, Li K, Duić N, Mi Z, Hodge BM, Shafie-khah M, et al. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. Energy Convers Manage 2018;171:839–54. https://doi.org/10.1016/j.enconman.2018.06.017.
[34] D'Oca S, Hong T. A data-mining approach to discover patterns of window opening and closing behavior in offices. Build Environ 2014;82:726–39. https://doi.org/10.1016/j.buildenv.2014.10.021.
[35] Zhang S, Zhang C, Yang Q. Data preparation for data mining. Appl Artificial Intellig 2003;17:375–81. https://doi.org/10.1080/713827180.
[36] Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Comput Stat Data Anal 2008;52:5186–201. https://doi.org/10.1016/j.csda.2007.11.008.
[37] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the 20th international conference on very large data bases, vol. 1215; 1994. p. 487–99. https://doi.org/10.1109/69.846291.
[38] Zaki MJ. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng 2000;12:372–90. https://doi.org/10.1109/69.846291.
[39] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: Proceedings of ACM SIGMOD international conference on management of data, vol. 29; 2000. p. 1–12. https://doi.org/10.1145/335191.335372.
[40] Rosenblatt M. Remarks on some nonparametric estimates of a density function. Ann Math Stat 1956;27(3):832. https://doi.org/10.1214/aoms/1177728190.
[41] Parzen E. On estimation of a probability density function and mode. Ann Math Stat 1962;33(3):1065. https://doi.org/10.1214/aoms/1177704472.
[42] Hipp J, Güntzer U, Nakhaeizadeh G. Algorithms for association rule mining—a general survey and comparison. ACM SIGKDD Explorat Newslett 2000;2:58–64. https://doi.org/10.1145/360402.360421.
[43] Deza MM, Deza E. Encyclopedia of distances. Berlin, Heidelberg: Springer; 2009https://doi.org/10.1007/978-3-642-00234-2_1.