Energy & Buildings 252 (2021) 111426

Contents lists available at ScienceDirect

Energy & Buildings

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

Analytic hierarchy process-based fuzzy post mining method for operation anomaly detection of building energy systems

Chaobo Zhang, Yang Zhao *, Jie Lu, Tingting Li, Xuejun Zhang

Institute of Refrigeration and Cryogenics, Zhejiang University, Hangzhou, China

A R T I C L E I N F O

Article history:
Received 18 May 2021
Revised 17 August 2021
Accepted 29 August 2021
Available online 1 September 2021

Keywords:
Energy efficiency enhancement for HVAC systems
Operation anomaly detection
Post mining
Association rule mining
Analytic hierarchy process
Fuzzy set theory

A B S T R A C T

Association rule mining has shown outstanding capacity in extracting operation patterns from numerous building operational data. However, only a very small portion of them is valuable for energy efficiency enhancement. Advanced post mining methods are necessary for automatically deleting most of worthless association rules. Therefore, this study proposes an analytic hierarchy process-based fuzzy post mining method. Three criteria and six corresponding sub-criteria are developed to evaluate the value of each association rule in energy efficiency enhancement. They offer new levels to assess the value of association rules. The fuzzy set theory is introduced to grade each sub-criterion of an association rule. It considers the uncertainties from the imprecise judgments, which can improve the robustness of rule assessing. Analytic hierarchy process is adopted to determine the weight of each criterion/sub-criterion for getting the weighted overall score of an association rule. It provides a solution to assessing the value of association rules from multiple aspects, which can improve the performance of post mining. The proposed method is evaluated using 117,636 association rules extracted from the one-year historical operational data of an actual chiller plant. Four common indexes (*support*, *confidence*, *lift* and distance correlation) are selected as a traditional method for comparison with the proposed method. The two methods can both filter out approximately 96.00% of worthless association rules. As for valuable association rules, the traditional method only extracts 17.28% of them, while this proportion is 93.51% for the proposed method. It proves that the proposed method has excellent performance of valuable association rule extraction.

1. Introduction

Buildings contribute to approximately 30% of the global energy consumption [1]. Heating, ventilation, and air conditioning (HVAC) systems are the largest energy consumers in buildings [2]. They consume about 50% of the energy consumption in buildings [3]. However, approximately 14% of the energy consumed by HVAC systems is wasted [4]. The waste usually results from various operation anomalies such as equipment performance degradation, inappropriate control strategies, and equipment/sensor faults [5]. It is becoming increasingly urgent to develop advanced energy efficiency enhancement technologies, such as optimal control [6–8] and fault detection/diagnosis [9–11], for HVAC systems.

Huge volumes of operational data of HVAC systems have been stored by building automation systems [12]. They can be analyzed for operation anomaly detection of HVAC systems. Manual analysis is usually difficult and time-consuming for such huge volumes of data. Data mining technologies can make the data analysis process more efficient and more convenient [13]. They are capable of extracting the operation anomalies of HVAC systems from their numerous operational data. Association rule mining (ARM) is one of the most promising data mining technologies in this field [14]. An association rule is usually expressed using “$A \rightarrow B$”, meaning “if A, then B”. It has shown powerful ability in discovering the interrelations among variables of high dimensions on a large data set. According to the previous studies, ARM can reveal the operation anomalies, such as sensor malfunctions, device malfunctions, and improper control, from numerous operational data of HVAC systems [15]. The mined association rules also can be further applied to establish expert systems for real-time fault detection and diagnosis of HVAC systems [16].

Yu et al. applied ARM to identify the improper control strategies and device malfunctions using the operational data of a variable air volume air-conditioning system for improving its energy efficiency [17]. With the help of the discovered knowledge, they developed an energy-efficient strategy for saving the energy in a heating coil. After that, they further developed an ARM-based framework for discovering the operation anomalies from numerous building-related data [18]. There were four components in this framework,

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

https://doi.org/10.1016/j.enbuild.2021.111426
0378-7788/© 2021 Elsevier B.V. All rights reserved.

Nomenclature

$Du_{\max}$	maximum duration of an association rule
Du_k	kth duration of an association rule
C	graph-based distance correlation index
D	minimum distance between a pair of variables
α	coefficient representing the physical similarity between a pair of variables
v	variable in an association rule
WS	device working state index
T_1	total time of devices related to an association rule simultaneously working
T_2	total time of an association rule occurring
R_S	Spearman correlation coefficient of an association rule
rg	rank variable
cov(.)	covariance of measurements between two rank variables
var(.)	variance of measurements of a rank variable
MIC	maximal information coefficient of an association rule
$I(X, Y)$	mutual information of the probability distributions of X and Y induced on the boxes of a grid
n_X	number of bins into which x-axis is partitioned
n_Y	number of bins into which y-axis is partitioned
$B(n)$	maximal grid size restriction function related to the sample size n
$\mathbf{A}$	pair-wise comparison matrix
a_{ij}	importance of the ith criteria/sub-criteria relative to the jth criteria/sub-criteria
$\lambda_{\max}$	largest eigenvalue of a pair-wise comparison matrix
ω	eigenvector
CI	consistency index of a pair-wise comparison matrix
CR	consistency ratio of the matrix
RI	random index of the matrix
$S_{overall}$	overall score of an association rule
$S_{i,j}$	score of the jth sub-criterion of the ith criterion
S_1	score of the stability criterion
S_2	score of the physical correlation criterion
S_3	score of the statistical correlation criterion
w_i	weight of the ith criterion
$w_{i,j}$	weight of the jth sub-criterion of the ith criterion
FR	flow rate
T	temperature
F	frequency
LR	load ratio
P	power
OOS	on–off state
TD	temperature difference
N	number of working devices

i.e., data import module, knowledge extraction module, knowledge output module and knowledge application module. Xiao and Fan also presented two similar ARM-based frameworks composed of four steps, i.e., data preprocessing, data partitioning, knowledge discovery and posting mining [19,20]. Data preprocessing aimed to clean raw data for obtaining high-quality data. Data partitioning was designed to partition the high-quality data into several subsets. In the step of knowledge discovery, ARM was adopted to extract knowledge from the data in different subsets. Posting mining was performed to choose, interpret and apply the knowledge. The two frameworks had been adopted to discover the improper control strategies [19], sensor faults [20], and common control strategies [21] from the operational data of HVAC systems. Based on the two frameworks, Li et al. proposed a similar framework including the same four steps for anomaly detection of variable refrigerant flow systems [22]. Device faults and improper control strategies were discovered using this framework from the operational data of a variable refrigerant flow system. Furthermore, Zhang et al. developed an improved ARM-based method for analyzing operational data of HVAC systems [23]. This method included three steps, i.e., kernel density estimation-based data preprocessing, ARM-based knowledge discovery and association rule comparison-based post mining. Improper control strategies and sensor malfunctions of a chiller plant were detected by this method. Qiu et al. chose the weighted ARM algorithm to identify HVAC systems' control strategies from observations with weights [24]. A weight represented the frequency of an observation occurring. On/off control strategies, sequencing control strategies and coordinated control strategies were discovered by this algorithm from the operational data of three buildings. Xue et al. detected the sensor faults and improper control strategies of a district heating substation using an ARM-based method [25]. Similar to the framework of Xiao and Fan [19,20], there were five steps in this method, i.e., data cleaning, data transformation, cluster analysis, ARM and knowledge interpretation. Improper control strategies of ground source heat pump systems also could be discovered by ARM according to [26]. Moreover, ARM has been adopted to explore the text data from HVAC systems. For instance, Gunay et al. proposed an ARM-based method to identify the device fault frequency using the text data from HVAC systems' maintenance work records [27]. Dutta et al. also applied ARM to assess occupant satisfaction using occupant survey data in the form of text data [28]. Temporal ARM is one specific ARM algorithm for discovering temporal rules which reveal one event will occur after the occurrence of another event. Fan et al. discovered that the power consumption of a chiller was high for some time using this algorithm [29]. Piscitelli et al. also applied this algorithm for an early fault detection of air handling units [30]. Gradual ARM is another specific ARM algorithm. A gradual rules indicate that the less/more one event is, the less/more another event is. This algorithm was utilized by Fan et al. to detect energy-inefficient operation patterns of a chiller plant such as heating and cooling counteraction [31].

The number of association rules is usually huge. But, most of them are worthless [32,33]. This issue is known as "post mining" in other industries such as insurance and retail [34]. The authors had investigated about one hundred thousand association rules mined from the operational data of a chiller plant [35]. Only 4.64% of them reveal the knowledge useful for energy efficiency enhancement. Therefore, it is necessary to develop post mining methods for removing the worthless association rules from HVAC systems automatically. The previous studies usually adopted three statistical indexes (*support*, *confidence*, and *lift*) for post mining. However, according to the authors' investigation [35], the three statistical indexes actually work badly in the domain of HVAC. They could not distinguish between valuable and worthless association rules well. To solve this problem, the authors proposed a graph-based correlation index for removing the association rules without physical meanings [36]. The results show that it can remove 90.11% of the worthless association rules. Nonetheless, this index cannot assess the value of the relationships between the variables that are not physically connected. Other aspects, such as statistical correlation, should be considered together with this index, so that the post mining process can be more generic and more powerful.

Analytic hierarchy process (AHP) is a multi-criteria decision-making method for helping decision makers choice suitable decisions [37,38]. It provides a solution to considering multiple aspects for assessing the value of an association rule. Based on this idea, this study proposes an AHP-based fuzzy post mining method. It grades the value of each association rule in improving the energy efficiency from three aspects (stability, physical correlation, and statistical correlation). The useful association rules can be identified based on their grades. Evaluation is made using the one-year historical operational data from an HVAC system' chiller plant. The main contortions of this study are summarized as follows:

- In this field, existing post mining methods usually assess the value of association rules using several indexes independently. An association rule will be regarded as worthless by such methods, if one index shows this rule is worthless, but another one indicates this rule is valuable. Such methods might make a wrong decision sometimes, since they cannot consider multiple aspects together. For instance, the results in Section 3.3.3 show that an existing method based on four common indexes mistakes 82.72% of valuable association rules for worthless association rules. With the aim of solving this problem, this work proposes a generic post mining framework based on AHP for assessing the value of association rules from three various aspects together. It can make the post mining process more reliable.
- Existing methods always use three statistical indexes (*support*, *confidence* and *lift*) for post mining of association rules. The three indexes have been demonstrated to be inefficient in this field [35]. Therefore, this study develops some new sub-criteria, such as maximum duration, frequency of occurrence, device working state and maximal information coefficient, to address this issue. They provide some new levels to evaluate the value of an association rule, which can improve the performance of post mining.
- AHP needs experts to determine the importance level of each criterion. Manual judgments are usually imprecise, which reduces the reliability of AHP. Therefore, this study introduces fuzzy set theory into AHP to take the uncertainties from the imprecise judgments into consideration. It can improve the robustness of value grading.

2. Methodology

The proposed AHP-based fuzzy post mining method includes three steps: establishing the hierarchical structure, determining the weights of criteria and sub-criteria, and calculating the overall scores of the mined association rules.

2.1. Establishing the hierarchical structure

This step aims to decompose the problem of association rule grading into a hierarchical structure. In general, a hierarchical structure can be visualized by a top-down hierarchical tree composed of several levels. The top level is the goal. The intermediate level is the criteria and sub-criteria. And the lowest level is the alternatives that need to be evaluated. Accordingly, a four-level hierarchical structure is proposed for association rule grading, as shown in Fig. 1. The goal is to evaluate the value of the mined association rules for improving the energy efficiency. The alternatives are all of the mined association rules. In this study, three types of criteria are proposed: stability, physical correlation, and statistical correlation. The three criteria are described in detail in the following sections.

2.1.1. First criterion: Stability

In general, transient operation patterns might be identified as operation anomalies wrongly. For example, the temperature of chilled water supplied by a chiller cannot drop down to the set point immediately when the chiller is just turned on. It might generate some misleading association rules, such as "the temperature of chilled water supplied by a chiller is 11 °C → the chiller is working". Such association rules seem like abnormal, because the temperature of chilled water supplied by a working chiller should be low. T However, they are worthless for operation anomaly identification, since they reveal transient operation patterns rather than operation anomaliesrather.

As shown in Fig. 2, the patterns hidden in an association rule generally occur many times. The duration of each occurrence should be very short if the association rule reveals a transient operation pattern. On the contrary, the duration of most occurrences should be long if the association rule reveals a stable operation pattern. Therefore, a sub-criterion, named maximum duration, is developed to quantify the stability of the patterns hidden in an association rule, as defined by Eq. (1). Its value will be large if it reveals stable operation patterns. And its value will be small if it reveals transient operation patterns.

$$Du_{\max} = \max_{k=1,2,\ldots,m} (Du_k) \tag{1}$$

where, $Du_{\max}$ is the maximum duration of the pattern hidden in an association rule, and Du_k is the kth duration of the pattern hidden in an association rule.

Some association rules might only occur several times. Such association rules might result from random events, such as sensors' signal fluctuations, regular device maintenance, and accidental mistakes resulting from manual control. They have very little influence on the energy efficiency of HVAC systems. Therefore, another sub-criterion, named frequency of occurrence, is presented. It is defined as the number of times the pattern hidden in an association rule occurs. The frequency of occurrence for the association rule shown in Fig. 2 is m, as it occurs m times.

2.1.2. Second criterion: Physical correlation

As reported in [35], most of the mined association rules do not have significant physical meanings. It is very crucial to remove the association rule without physical meanings. In this study, two sub-criteria are presented to detect such association rules, including a graph-based distance correlation index, and a device working state index.

If a pair of variables are monitored from the sensors or devices that are close to each other, they are generally more possible to be physically correlative [36]. Moreover, variables should be valuable for revealing HVAC systems' coordinated control strategies if they are monitored from the same type of device (such as chillers, pumps, and cooling towers), and have the same type of meaning (such as temperature, power, and frequency) [36]. For instance, two chillers should have similar supply chilled water temperatures if they work simultaneously. According to the above two concepts, a graph-based distance correlation index (C) is presented as shown in Eq. (2). It can quantify the physical correlation of an association rule. The minimum distance in Eq. (2) is the shortest path between two variables in an undirected graph representing the topological structure of an HVAC system. The undirected graph is named as "variable network graph". Three types of nodes are included in a variable network graph, i.e., device nodes, sensor nodes, and virtual variable nodes. Device nodes and sensor nodes represent the variables from devices and sensors, respectively. Virtual variable nodes

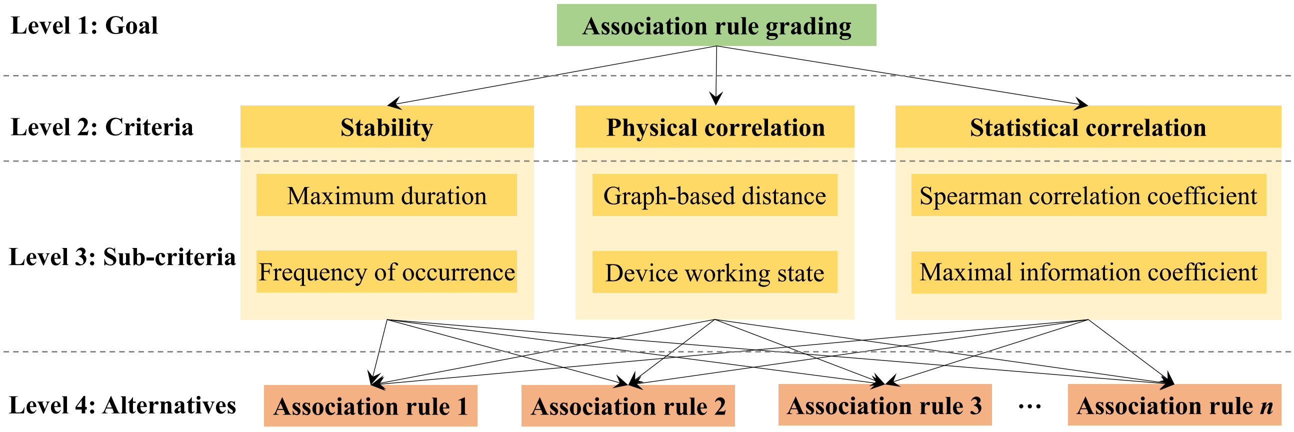

Fig. 1. Four-level hierarchical structure for association rule grading.

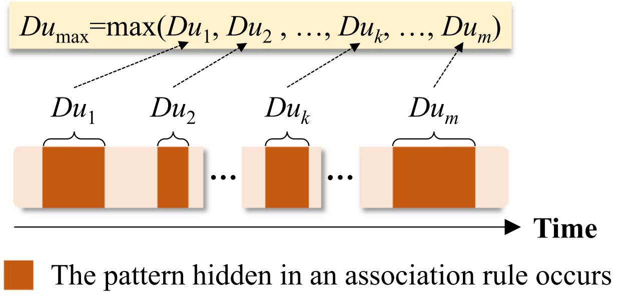
Du_{max}=max(Du_1, Du_2 , …, Du_k, …, Du_m)
Du_1
Du_2
Du_k
Du_m
…
…
Time
The pattern hidden in an association rule occurs

Fig. 2. Illustration of the definition of the duration of an association rule.

represent the variables that cannot be measured but can be calculated based on known variables. Virtual variable nodes, such as the temperature difference of chilled water, and number of working devices, are useful for detecting the operation anomalies. Two nodes are connected by an edge with a weight according to the topological structure of an HVAC system. The weight of two variables in an edge represents their physical distance. Weights are determined based on five principles (see Table 1), as suggested in [36]. The weight of an edge should be relatively large if the edge is related to devices, as devices influence the medium's state parameters significantly but sensors do not. The minimum distance between a pair of variables is the minimum sum of weights of a path between them. It is a shortest path problem that can be solved by many algorithms, such as Dijkstra [39], Bellman [40], and Floyd-Warshall [41]. A hypothetical HVAC system (see Fig. 3) is taken as an example to clarify how to construct a variable network graph (see Fig. 4).

$$C = \max_{v_i \neq v_j}\left(\alpha_{v_i,v_j} \times D_{v_i,v_j}\right) \tag{2}$$

where, C is the graph-based distance correlation of an association rule, α is a coefficient representing whether two diverse variables are monitored from the same type of device and have the same type of meaning, D is the minimum distance between two variables, and v_i and v_j are two diverse variables in the association rule. As suggested in [36], the value of α can be set as 0.5 if two variables are monitored from the same type of device and have the same type of meaning. Otherwise, the value of α can be set as 1.0.

Table 1
Principles for determining the weight of an edge in a variable network graph.

No.	Description	Weight
1	Two variables on the edge are related to two diverse devices.	1.0
2	Two variables on the edge are related to a sensor and a device, respectively.	0.5
3	One variable on the edge is related to a virtual variable node.	0.5
4	Two variables on the edge are related to sensors.	0.0
5	Two variables on the edge are related to the same device.	0.0

In general, the operation patterns of non-working devices are worthless for revealing the operation anomalies. Therefore, a device working state index is developed to identify the association rules related to non-working devices, as defined by Eq. (3). The working state of a device can be identified based on its power or on–off state. For instance, a device can be regarded as always working if more than 95% of its on–off state measurements are "on". An association rule might be related to more than one device. All the related devices are considered in the device working state index.

$$WS = \frac{T_1}{T_2} \tag{3}$$

where, WS is the device working state index of an association rule, T_1 is the total time devices related to the association rule take to simultaneously work when the patterns hidden in the association rule occurs, and T_2 is the total time of the patterns hidden in the association rule occurring.

2.1.3. Third criterion: Statistical correlation

Apart from the physical correlation, statistical correlation is also crucial for discovering the potential relations between variables. In this study, the Spearman correlation coefficient and maximal information coefficient are selected as two sub-criteria to discover the statistical correlation between variables. Some variables might be unrelated when devices are not working. For example, the supply chilled water temperature of a non-working chiller should be unrelated to that of another working chiller. Therefore, for such variables, only measurements collected during the working time of devices are utilized to estimate the statistical correlation between them.

The Spearman correlation coefficient is a common statistical index for measuring the linear relations and monotonic nonlinear relations between two variables [42]. It is Pearson correlation coefficient between rank variables in essence, but it has higher robustness than Pearson correlation coefficient. Rank variables are defined as the rank values of the measurements of raw numerical variables. Spearman correlation coefficients range from −1 (perfectly negative correlation) to + 1 (perfectly positive correlation). Considering an association rule might be related to more than

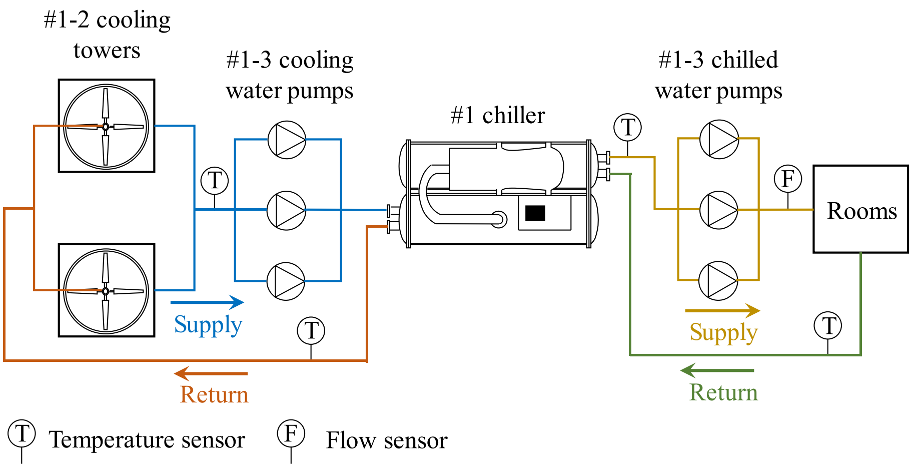

Fig. 3. Illustration of a hypothetical HVAC system.

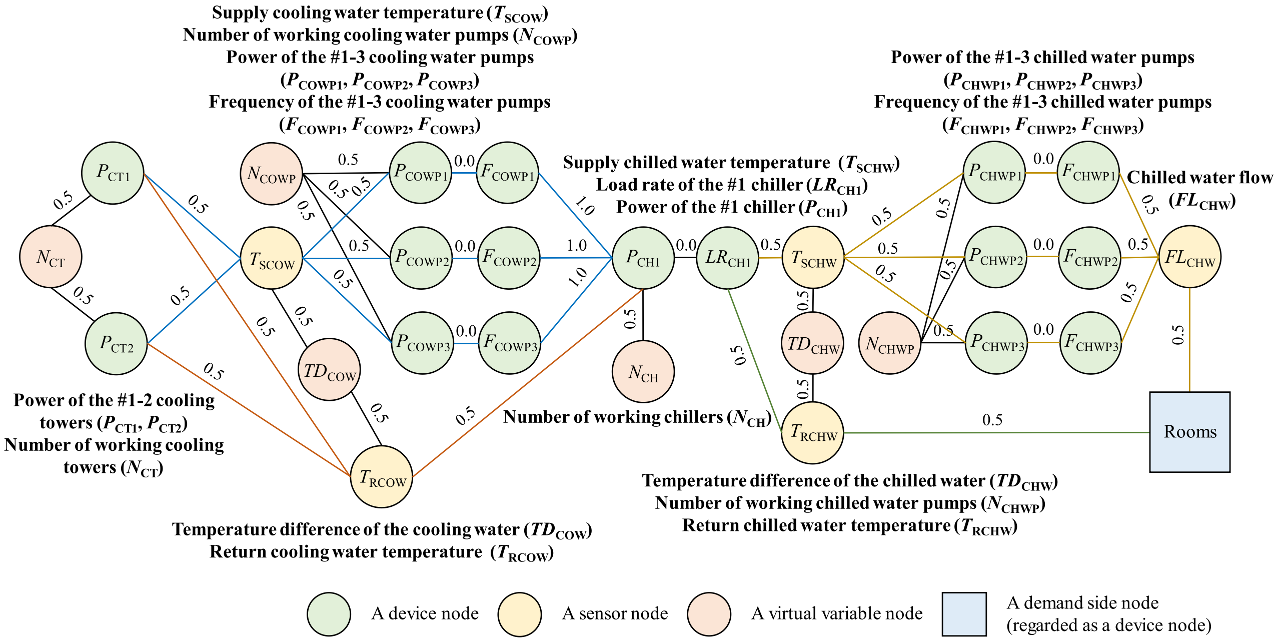
Supply cooling water temperature (T_{SCOW})
Number of working cooling water pumps (N_{COWP})
Power of the #1-3 cooling water pumps (P_{COWP1}, P_{COWP2}, P_{COWP3})
Frequency of the #1-3 cooling water pumps (F_{COWP1}, F_{COWP2}, F_{COWP3})
Power of the #1-3 chilled water pumps (P_{CHWP1}, P_{CHWP2}, P_{CHWP3})
Frequency of the #1-3 chilled water pumps (F_{CHWP1}, F_{CHWP2}, F_{CHWP3})
Supply chilled water temperature (T_{SCHW})
Load rate of the #1 chiller (LR_{CH1})
Power of the #1 chiller (P_{CH1})
Chilled water flow (FL_{CHW})
P_{CT1}
N_{CT}
P_{CT2}
N_{COWP}
T_{SCOW}
TD_{COW}
T_{RCOW}
P_{COWP1}
P_{COWP2}
P_{COWP3}
F_{COWP1}
F_{COWP2}
F_{COWP3}
P_{CH1}
N_{CH}
LR_{CH1}
T_{SCHW}
TD_{CHW}
T_{RCHW}
N_{CHWP}
P_{CHWP1}
P_{CHWP2}
P_{CHWP3}
F_{CHWP1}
F_{CHWP2}
F_{CHWP3}
FL_{CHW}
Rooms
0.5
0.0
1.0
Power of the #1-2 cooling towers (P_{CT1}, P_{CT2})
Number of working cooling towers (N_{CT})
Number of working chillers (N_{CH})
Temperature difference of the cooling water (TD_{COW})
Return cooling water temperature (T_{RCOW})
Temperature difference of the chilled water (TD_{CHW})
Number of working chilled water pumps (N_{CHWP})
Return chilled water temperature (T_{RCHW})
A device node
A sensor node
A virtual variable node
A demand side node (regarded as a device node)

Fig. 4. Illustration of the hypothetical HVAC system's variable network graph.

two variables, the Spearman correlation coefficient of an association rule is represented as the minimum of absolute values of Spearman correlation coefficients of every pair of different rank variables in the association rule, as shown in Eq. (4).

$$R_S = \min_{rg_i \neq rg_j} \left(\text{abs} \left(\frac{cov(rg_i, rg_j)}{\sqrt{var(rg_i) \times var(rg_j)}} \right) \right) \tag{4}$$

where, R_S is the Spearman correlation coefficient of an association rule, rg_i and rg_j are two different rank variables in the association rules, cov(.) is the covariance of measurements of two rank variables, and var(.) is the variance of measurements of a rank variable.

Considering the Spearman correlation coefficient cannot quantify the complicated non-linear relations between two variables, the maximal information coefficient is also applied as another sub-criterion. It is a mutual information-based nonparametric statistical index proposed by Reshef et al. [43]. It can not only capture linear and non-linear correlations but also measure relations without a specific function type [44]. It ranges from 0.0 and 1.0. The higher the value of the maximal information coefficient between two variables is, the more correlative the two variables are. Two variables are independent if the value of the maximal information coefficient between them is 0.0. An association rule might contain more than two variables. Therefore, the maximal information coefficient of an association rule is represented as the minimum of maximal information coefficients of every pair of various variables in the association rule, as shown in Eq. (5).

$$MIC = \min_{X \neq Y} \left(\max_{n_X \times n_Y < B(n)} \left(\frac{I(X, Y)}{\log_2(\min(n_X, n_Y))} \right) \right) \tag{5}$$

where, MIC is the maximal information coefficient of an association rule, X and Y are two various variables in the association rule, $I(X, Y)$ is the mutual information of the probability distribution of X and Y induced on the boxes of a grid, n_X is the number of bins into which x-axis is partitioned, n_Y is the number of bins into which y-axis is partitioned, and $B(n)$ is the maximal grid size restriction function related to the sample size n. In general, $B(n)$ is suggested to be $n^{0.6}$ [43].

2.2. Determining the weights of criteria and sub-criteria

Based on the established hierarchy structure, three sub-steps are adopted to determine the weights of criteria and sub-criteria. Firstly, pair-wise comparison matrices are constructed to quantify the relative importance between each two criteria and each two sub-criteria. The pair-wise comparison scale proposed by Saaty [45] is utilized, as listed in Table 2. A pair-wise comparison matrix is defined by Eq. (6). The value of a_{ij} is equal to the reciprocal of a_{ji} in the pair-wise comparison matrix.

$$\boldsymbol{A} = \begin{bmatrix} 1 & a_{12} & \dots & a_{1n} \\ 1/a_{12} & 1 & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ 1/a_{1n} & 1/a_{2n} & \dots & 1 \end{bmatrix} \tag{6}$$

where, $\mathbf{A}$ is a pair-wise comparison matrix, and a_{ij} is the importance of the ith criteria/sub-criteria relative to the jth criteria/sub-criteria.

Secondly, the largest eigenvalue of a pair-wise comparison matrix and its corresponding eigenvector are calculated. The relation among the largest eigenvalue, eigenvector and pair-wise comparison matrix is defined by Eq. (7). The eigenvector is then normalized to obtain the weight vector of corresponding criteria/sub-criteria.

$$\boldsymbol{A\omega} = \lambda_{\max}.\boldsymbol{\omega} \tag{7}$$

where, $\mathbf{A}$ is a pair-wise comparison matrix, $\lambda_{\max}$ is the largest eigenvalue of the pair-wise comparison matrix, and ω is the corresponding eigenvector.

Thirdly, the consistency of a pair-wise comparison matrix is evaluated, as the inconsistency may happen due to the subjective judgment. Two common indexes, named consistency index and consistency ratio, are adopted to evaluate the consistency [46]. They are defined by Eqs. (8) and (9), respectively. A pair-wise comparison matrix is regarded as consistent if the value of CR is less than 0.1, or the value of CI is equal to 0.0. Otherwise, the pair-wise comparison matrix should be constructed again until it is consistent.

$$CI = \frac{\lambda_{\max} - n}{n - 1} \tag{8}$$

$$CR = \frac{CI}{RI} \tag{9}$$

where, CI is the consistency index of a pair-wise comparison matrix, CR is the consistency ratio of the matrix, RI is the random index of the matrix, $\lambda_{\max}$ is the largest eigenvalue of the matrix, and n is number of criteria/sub-criteria in the matrix. The value of RI is determined according to Table 3 [46].

Table 2
Pair-wise comparison scale of criteria and sub-criteria.

Scale of importance	Definition	Explanation
1	Equal importance	Two criteria/sub-criteria contribute equally to the goal.
3	Moderate importance	A criteria/sub-criteria is favored slightly over another.
5	Strong importance	A criteria/sub-criteria is favored strongly over another.
7	Very strong importance	A criteria/sub-criteria is favored very strongly over another.
9	Extreme importance	A criteria/sub-criteria is favored extremely over another.
2, 4, 6, 8	Intermediate values between two adjacent scale values	Make a compromise between two adjacent judgments.

2.3. Calculating the overall scores of association rules

The last step aims to calculate the overall scores of the mined association rules for quantifying the value of the mined association rules in improving the energy efficiency. Before calculating the overall score, it is necessary to calculate the score of each sub-criterion. A five-level scoring standard is adopted in this study, as listed in Table 4. Then, the overall score of an association rule is calculated using Eq. (10). The higher the overall score of an association rule is, the more valuable the association rule is.

$$S_{overall} = \sum_{i=1}^{p} \left(w_i \times \sum_{j=1}^{q_i} \left(w_{i,j} \times S_{i,j} \right) \right) \tag{10}$$

where, $S_{overall}$ is the overall score of an association rule, w_i is the weight of the ith criterion, $w_{i,j}$ is the weight of the jth sub-criterion of the ith criterion, $S_{i,j}$ is the score of the jth sub-criterion of the ith criterion, p is the number of criteria, and q_i *is the number of sub-criteria of the ith criterion*.

Considering the uncertainties caused by imprecise judgments, fuzzy sets are established for providing a mathematical representation to describe the vagueness and fuzziness using an index named degree of membership [47,48]. The degree of membership ranges from 0.0 to 1.0, indicating the degree of uncertainty that a value belongs to a given set. The value does not belong to the given set if its degree of membership is 0.0. And the value completely belongs to the set if its degree of membership is 1.0. Taking the maximum duration as an example, an association rule is weakly valuable if its maximum duration is very short, as it usually reflects the transient operation patterns. Therefore, the score of the maximum duration should be small for an association rule whose maximum duration is very short. This vague description can be represented by fuzzy sets shown Fig. 5. Based on the fuzzy sets, a specific value of a sub-criterion is determined as belonging to the fuzzy set with the maximum degree of membership.

3. Evaluation

3.1. Description of the evaluation

The performance of the AHP-based fuzzy post mining method is valuated in the way shown in Fig. 6. It includes four steps: data preprocessing, knowledge discovery, knowledge post mining, and knowledge application. A representative chiller plant serving a public building in Shenzhen, China, is chosen as the data source to validate the performance of the proposed post mining method. This chiller plant consists of twenty cooling towers (CT1-CT20), ten cooling water pumps (COWP1-COWP10), eight chillers (CH1-CH8), fourteen secondary chilled water pumps (SCHWP1-SCHWP14), and ten primary chilled water pumps (PCHWP1-PCHWP10), as shown in Fig. 7. The four types of devices are the most common components for most chiller plants. Abundant sensors are installed in this chiller plant. They monitor a total of 148 variables, including the outdoor meteorological parameters (outdoor air temperature, and outdoor air relative humidity), operating parameters of devices and valves (power, frequency, on–off states, and load ratio), and state parameters of chilled and cooling water (temperature, and flow rate). Based on these monitored variables, 6 virtual variables are further calculated, including the number of working cooling towers, number of working cooling water pumps, number of working chillers, number of working primary chilled water pumps, temperature difference of chilled water, and temper-

Table 3
Value of *RI* for matrices of order 1 to 5.

Matrix order	1	2	3	4	5
RI	0.00	0.00	0.58	0.90	1.12

Table 4
Scoring standard for sub-criteria of an association rule.

Score	Definition
1	An association rule is weakly valuable.
3	An association rule is moderately valuable.
5	An association rule is strongly valuable.
2, 4	Intermediate values between two adjacent scores.

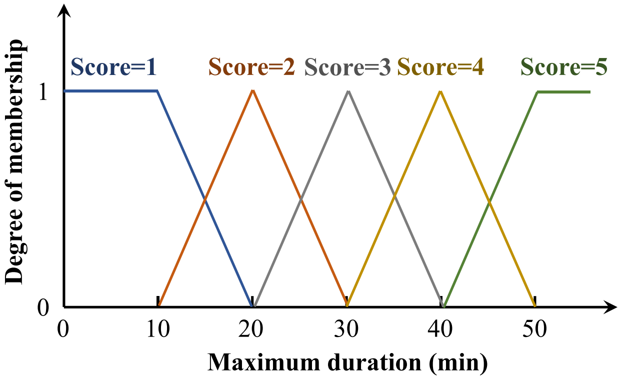

Fig. 5. Membership functions of fuzzy sets for the maximum duration.

ature difference of cooling water. Finally, the historical operational data of the 154 variables collected in 2016 with a sampling interval of 10 min are used for evaluation. It needs to be noted that only a chiller plant system is utilized for evaluation, since the proposed method doesn't consider the difference in working characteristics and system structure of a chiller plant system. It should be applicable to other types of chiller plant systems, if it works well on a representative chiller plant system.

3.2. Data preprocessing and knowledge discovery

3.2.1. Results of data preprocessing

A kernel density estimation-based statistical approach is adopted for numerical variables' outlier detection [23]. It calculates the probability density function using the measurements of a numerical variable. Measurements are detected as outliers if their probabilities are lower than a threshold. Outliers and missing values should be removed if they last for a long time. Otherwise, they can be substituted by the values calculated by some statistical algorithms, such as linear interpolation. As suggested in [19], missing values and outliers are substituted by the values calculated using linear interpolation in this study if they last for less than one h. They are removed if they last for more than one h. Finally, 0.77% of the raw historical operational data are removed.

A common data transformation approach, named equal-width binning, is then utilized to transform numerical data into categorical data. The bin widths are 5.0 °C, 5.0 °C, 2.0 °C, 10.0%, 100.0 m^3/h, 2.0 Hz, and 1.0 K for the outdoor air temperature, cooling water temperature, chilled water temperature, outdoor air relative humidity, flow rate, pump frequency, and temperature difference, respectively. For the load ratio and power, their measurements are classified into four classes, i.e., *low*, *medium*, *high*, and *off*. The class of *off* indicates a device is non-working. The classes of *low*, *medium*, and *high* indicate the load ratio or power of a device is low, medium, and high, respectively. The bin widths in the three categories are equal to (*maximum* - *minimum*)/3.

3.2.2. Results of knowledge discovery

Association rules are then mined by the FP-growth algorithm [23]. It needs to be noted that *support*, *confidence*, and *lift* are not utilized to restrain the generation of association rules in this study, because the authors discovered that they were actually ineffective in this field [35]. Association rules are represented using "$A \rightarrow B$", indicating "if *A*, then *B*". Because both "$B \rightarrow A$" and "$A \rightarrow B$" are available, a merged association rule named bidirectional association rules ("$A \leftrightarrow B$") is used. Only association rules related to two variables are utilized in this study, as the number of association rules will be too large to analyze if there are more than two variables in an association rule. Finally, 117,636 bidirectional association rules are extracted.

3.3. Evaluation of the AHP-based fuzzy post mining method

3.3.1. Weights of the criteria and sub-criteria

The hierarchical structure shown in Fig. 1 is utilized to grade the value of the 117,636 bidirectional association rules. The weights of the criteria and sub-criteria in the hierarchical structure are first determined. Fig. 8 illustrates the ideal sorting of scores of association rules in our opinions. Stable rules should have higher scores than unstable rules. And stable rules with high physical correlations should have higher scores than those with high statistical correlations, as a rule related to non-working devices (weak physical correlations) should be worthless even if the variables in the rule are highly correlative in statistics. Based on the ideal sorting, the relative importance between each two criteria are determined by one expert according to the pair-wise comparison scale shown in Table 2. A comparison matrix is then constructed to determine the weight of each criterion, as listed in Table 5. The weights are 0.63, 0.24, and 0.14 for the stability, physical correlation, and statistical correlation, respectively. Similarly, three pair-wise comparison matrixes are further constructed by the expert to calculate the weights of the sub-criteria, as listed in Tables 6–8. The weights of the sub-criteria of statistical correlation are both 0.50, because it is assumed that there is no preference between them. For the stability, the weight of the maximum duration is higher than that of the frequency of occurrence, because an association rule related to transient operation patterns should be worthless even if it might occur many times. For the physical correlation, the weight of the device working state is higher than that of the graph-based distance, because a rule related to non-working devices should be worthless even if the variables in the rule are physically close to each other. In this study, the pair-wise comparison matrixes are constructed by one expert. If there are many experts, some group decision making methods can be adopted to aggregate the pair-wise comparison matrixes of different experts [49].

3.3.2. Fuzzy sets of the sub-criteria

After determining the weights of the criteria and sub-criteria, the fuzzy sets of the sub-criteria are further established to assign a certain value of a sub-criterion to a certain score using a vague description. The membership functions of the fuzzy sets of the sub-criteria

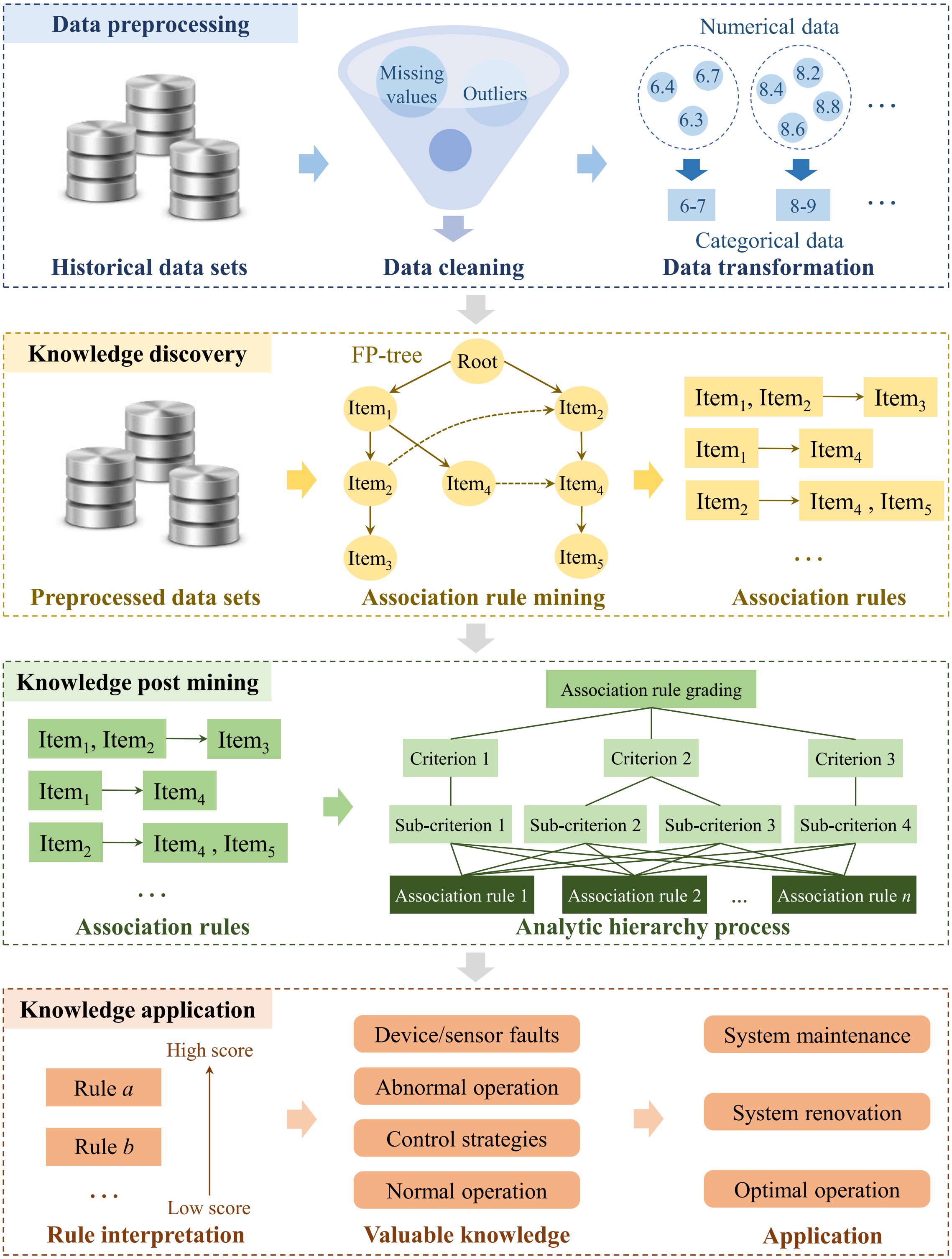

Fig. 6. Evaluation process of the AHP-based fuzzy post mining method.

are shown in Fig. 9. Based on the graph-based distance correlation index's definition, the smaller the graph-based distance of an association rule is, the more possible the variables in the rule are to be correlative. Therefore, the score of an association rule is high if its graph-based distance is small, as shown in Fig. 9 (c). For the other five sub-criteria, the larger the values of these sub-criteria of an association rule are, the more possible the rule are to be valuable for energy efficiency enhancement. Hence, the score of an association rule is high if the values of these sub-criteria are large.

3.3.3. Grading the value of the mined association rules

The values of the sub-criteria of each association rule are then calculated to determine the scores of the sub-criteria of each association rule, according to the fuzzy sets of the sub-criteria. The

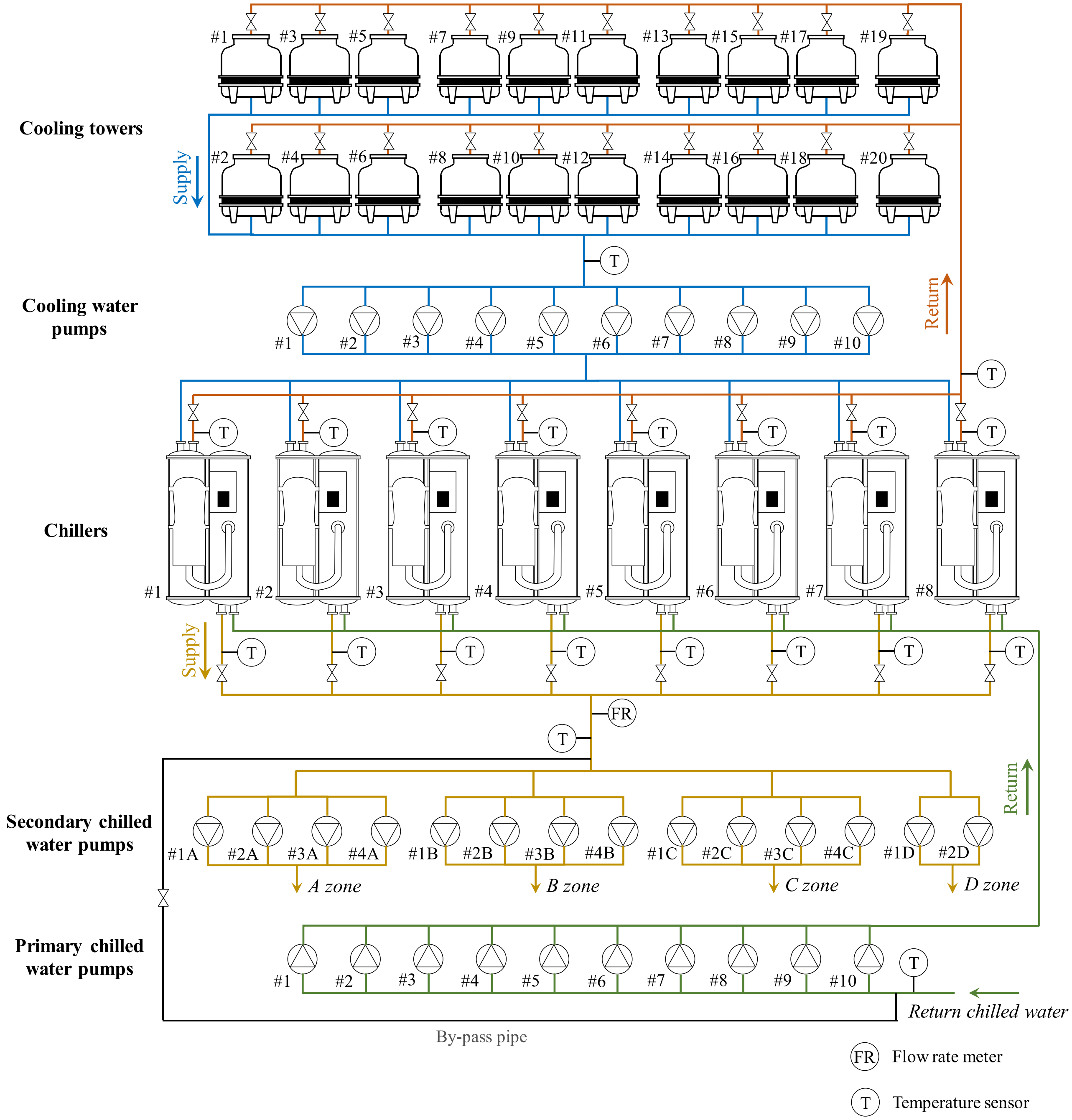

Fig. 7. Illustration of the chiller plant.

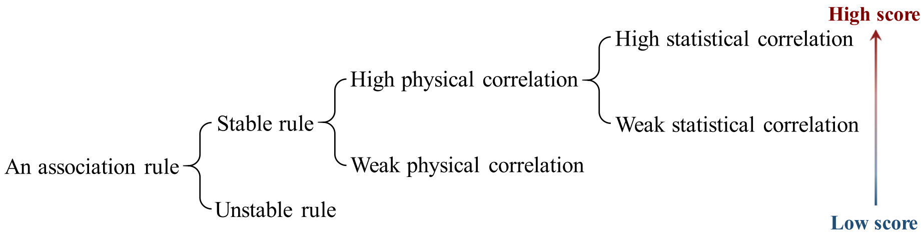
High score
High statistical correlation
High physical correlation
Stable rule
Weak statistical correlation
An association rule
Weak physical correlation
Unstable rule
Low score

Fig. 8. Ideal sorting of scores of association rules.

Floyd-Warshall algorithm is adopted to compute the minimum weighted distance between two variables, due to its simplicity [50]. A variable network graph is constructed to compute the graph-based distance, as shown in Fig. 10. The outdoor air temperature and outdoor air relative humidity cannot be included in the variable network graph, as they are not physically connected to

Table 5
Pair-wise comparison matrix of the three criteria (CR = 0.02).

	Stability	Physical correlation	Statistical correlation	Weights
Stability	1	3	4	0.63
Physical correlation	1/3	1	2	0.24
Statistical correlation	1/4	1/2	1	0.13

Table 6
Pair-wise comparison matrix of the two sub-criteria of the stability (CI = 0.00).

	Maximum duration	Frequency of occurrence	Weights
Maximum duration	1	2	0.67
Frequency of occurrence	1/2	1	0.33

Table 7
Pair-wise comparison matrix of the two sub-criteria of the physical correlation (CI = 0.00).

	Graph-based distance	Device working state	Weights
Graph-based distance	1	1/2	0.33
Device working state	2	1	0.67

Table 8
Pair-wise comparison matrix of the two sub-criteria of the statistical correlation (CI = 0.00).

	Spearman correlation coefficient	Maximal information coefficient	Weights
Spearman correlation coefficient	1	1	0.50
Maximal information coefficient	1	1	0.50

other variables. Therefore, the graph-based distance of association rules related to the two variables cannot be calculated. For such association rules, the scores of their graph-based distance are set as 3 (a moderate value). They still will get relatively high scores if they have high statistical correlations.

Finally, the overall score of each association rule is calculated using Eq. (10) based on the scores and weights of its sub-criteria. With the aim of revealing the relations among the score of stability (S_1), score of physical correlation (S_2), score of statistical correlation (S_3), and overall score (S_{overall}), the k-means clustering algorithm is utilized to classify them based on their Euclidean distance. The Calinski-Harabasz index is adopted to optimize the number of clusters [51]. The higher the value of this index is, the better the quality of classified clusters is. The number of clusters changes from 2 to 20 with an interval of 1. The value of the Calinski-Harabasz index is the highest when the number of clusters is 9. The classified clusters are visualized in Fig. 11 using violin plots. The first cluster shows that the overall score is very high when the scores of all the three criteria are high. It is reasonable because such association rules are stable and have significant physical and statistical correlations. The second and third clusters reveal that the overall score is relatively high when the score of stability and one of the scores of physical and statistical correlations are both high. Such association rules are stable and have significant physical or statistical correlations, which also might be useful. The fourth cluster shows that the overall score is relatively low when the score of stability is high but the scores of physical and statistical correlations are both low. Such association rules are stable but the variables in the rules are always not correlative. Therefore, they are usually worthless. Other clusters show that the overall score is relatively low when the score of stability is relatively low. It is reasonable because unstable association rules are usually related to transient operation patterns.

The association rules are ranked in the order of their overall scores. They are checked orderly according to our engineering experience in data science and building energy conservation for obtaining the distributions of overall scores of worthless and valuable association rules, respectively. An association rule is labeled as valuable if it reveals a stable operation pattern that can inspire us to make a decision for energy efficiency enhancement. And an association rule is labeled as worthless if it reveals a transient operation pattern, or it indicates a stable operation pattern that cannot inspire us to make a decision for energy efficiency enhancement. A total of 1,418 valuable association rules and 116,218 worthless association rules are identified. Based on the labeled association rules, the probability density distributions and cumulative probability density distributions of overall scores of worthless and valuable association rules are shown in Figs. 12 and 13, respectively. As shown in Fig. 12, the overall scores of most valuable association rules are significantly higher than those of most worthless association rules. According to Fig. 11, the medians of S_{overall} in the first three clusters are 4.50, 4.30, and 4.00, respectively. Therefore, the threshold of S_{overall} is suggested to be within the range of 4.0 to 4.5. Fig. 13 indicates that the higher this threshold is, the more worthless association rules are removed. However, this threshold cannot be too high, as more valuable association rules will be removed wrongly with the increase of this threshold. The threshold is set to 4.35 in this study to ensure that most of the worthless association rules can be removed and only a small portion of the valuable association rules are removed wrongly. As shown in Fig. 13, S_{overall} of 95.45% of the worthless association rules are lower than this threshold, and S_{overall} of 93.51% of the valuable association rules are higher than this threshold. It proves that the AHP-based fuzzy post mining method can isolate the valuable association rules from the worthless association rules effectively. Users just need to analyze the association rules with high overall scores. It can significantly improve the efficiency of knowledge discovery in practice.

In previous studies, three statistical indexes (*support*, *confidence* and *lift*) [17–22] and the graph-based distance correlation index (C) [36] have been utilized for post mining of association rules. They are adopted as a traditional method for performance comparison with the proposed method. As suggested in [36], the minimum thresholds of *support* and *confidence* are set to be 5.00% and 60.00%, respectively. The minimum threshold of *lift* is set to be 1.00 as recommended in [17–22]. The maximum threshold of C is set to be 0.50, as most worthless association rules have a C higher than 0.5 in this study. This traditional method removes 96.32% of the worthless association rules. However, it only extracts 17.28% of the valuable association rules. Although the proposed method and the traditional method can both filter out most of the worthless association rules, the traditional method deletes many valu-

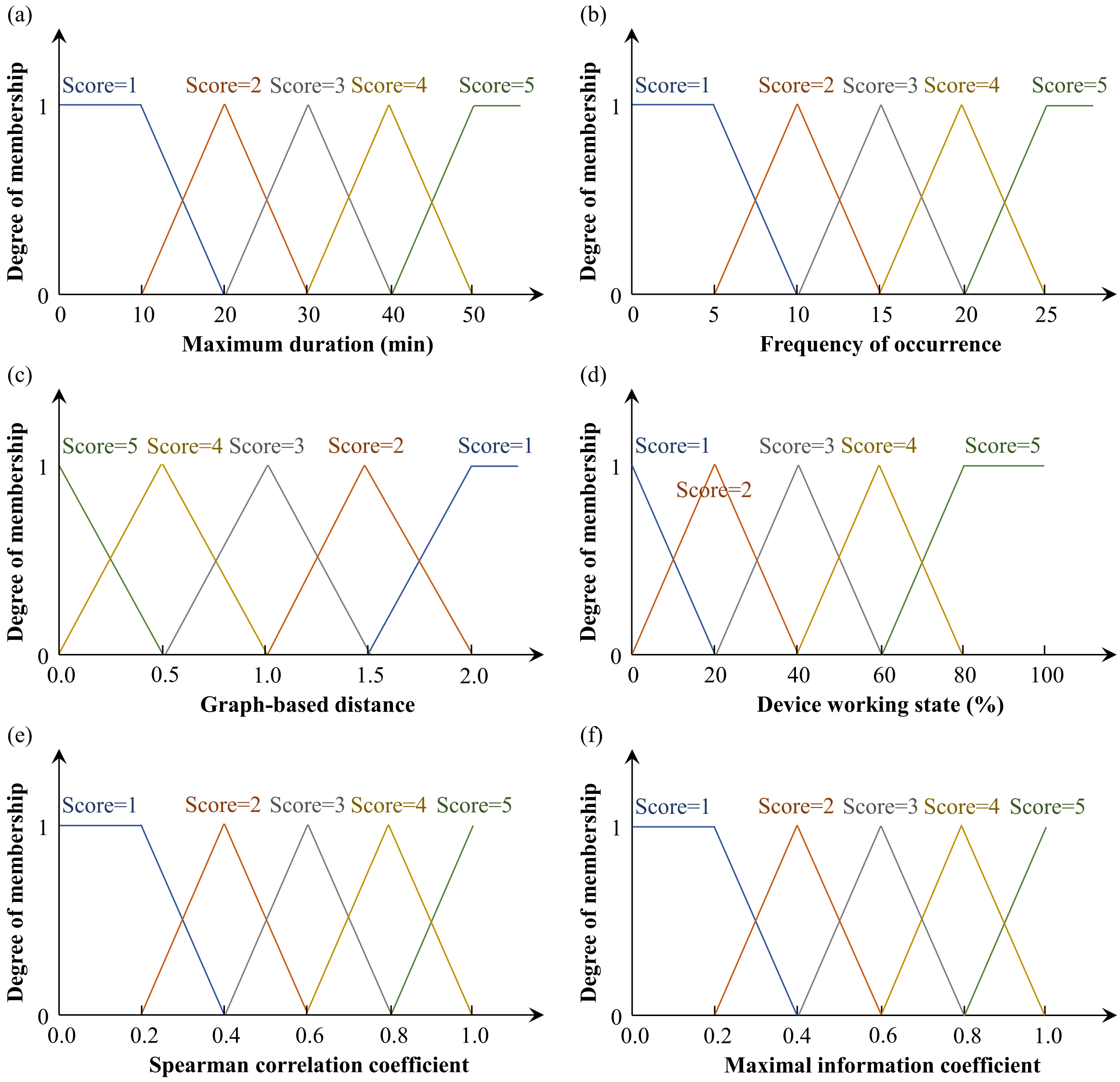

Fig. 9. Membership functions of the fuzzy sets of the sub-criteria.

able association rules (82.72%) wrongly. The main reason is that the traditional method assesses an association rule based on the four indexes independently. A valuable association rule will be mistaken for a worthless association rule if one index indicates the rule is valuable but another one doesn't. It proves that the proposed method has better performance of extracting valuable association rules.

3.3.4. Typical knowledge discovered for improving the energy efficiency

3.3.4.1. Energy-inefficient control of parallel chilled water pumps. Thirteen bidirectional association rules associated with the frequency of #6 primary chilled water pump and frequency of #7 primary chilled water pump are discovered, as shown in Fig. 14. It is discovered that the frequency of #6 primary chilled water pump was different from the frequency of #7 primary chilled water pump sometimes in this chiller plant. Moreover, other primary chilled water pumps also existed the same issue. It results in the energy waste, as it is generally the most efficient that parallel variable-frequency pumps operate at the same frequency [52]. It is necessary to optimize the coordinated control strategy of the parallel chilled water pumps in this chiller plant to ensure that they always operate at the same frequency.

3.3.4.2. Faulty chilled water valve in a chiller. Ten bidirectional association rules related to #7 chiller are extracted, as shown in Fig. 15. It is found that the #7 chiller's chilled water valve was turned off sometimes when #7 chiller was working. The #7 chiller's supply chilled water temperature was very low (between 2 °C and 6 °C) when this fault occurred. It was below the normal interval (between 6 °C and 10 °C) of the #7 chiller's supply chilled water temperature when the chilled water valve was fault-free. This fault resulted in the energy waste in #7 chiller. It even would threaten the safe operations of #7 chiller, as the chilled water might freeze. It is necessary to adopt some interlocking control strategies to avoid such faults. For instance, a chiller cannot be turned on when its chilled water valve is closed.

3.3.4.3. Energy-inefficient operations of chillers under low cooling energy demand. Six bidirectional association rules related to the energy-inefficient operations of chillers under low cooling energy demand are found, as shown in Fig. 16. It is discovered that #8 chil-

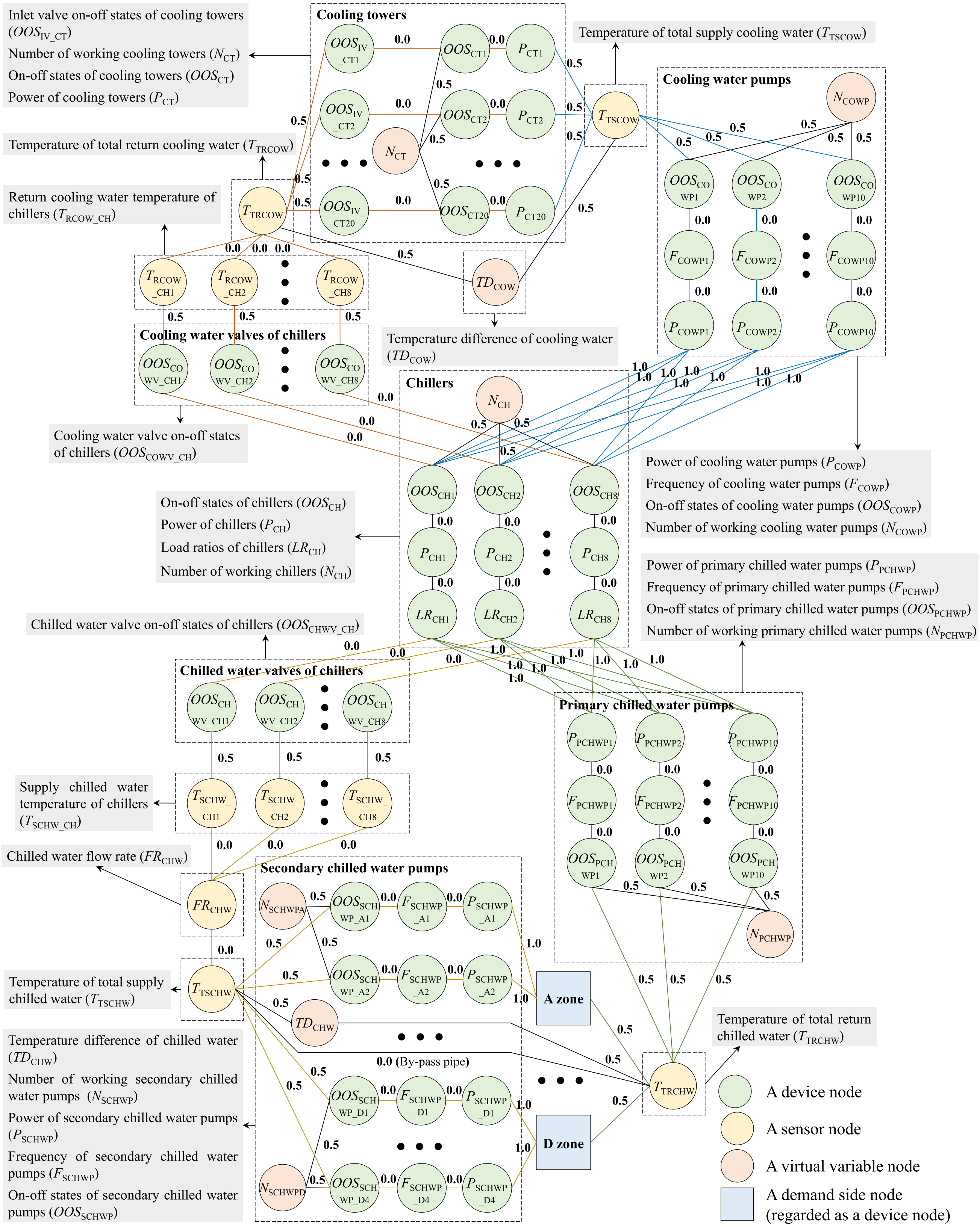
Inlet valve on-off states of cooling towers (OOS_{IV_CT})
Number of working cooling towers (N_{CT})
On-off states of cooling towers (OOS_{CT})
Power of cooling towers (P_{CT})
Cooling towers
OOS_{IV_CT1}
OOS_{IV_CT2}
OOS_{IV_CT20}
OOS_{CT1}
OOS_{CT2}
OOS_{CT20}
P_{CT1}
P_{CT2}
P_{CT20}
N_{CT}
Temperature of total supply cooling water (T_{TSCOW})
T_{TSCOW}
Temperature of total return cooling water (T_{TRCOW})
T_{TRCOW}
Return cooling water temperature of chillers (T_{RCOW_CH})
T_{RCOW_CH1}
T_{RCOW_CH2}
T_{RCOW_CH8}
TD_{COW}
Temperature difference of cooling water (TD_{COW})
Cooling water pumps
N_{COWP}
OOS_{COWP1}
OOS_{COWP2}
OOS_{COWP10}
F_{COWP1}
F_{COWP2}
F_{COWP10}
P_{COWP1}
P_{COWP2}
P_{COWP10}
Cooling water valves of chillers
OOS_{COWV_CH1}
OOS_{COWV_CH2}
OOS_{COWV_CH8}
Cooling water valve on-off states of chillers (OOS_{COWV_CH})
Chillers
N_{CH}
OOS_{CH1}
OOS_{CH2}
OOS_{CH8}
P_{CH1}
P_{CH2}
P_{CH8}
LR_{CH1}
LR_{CH2}
LR_{CH8}
Power of cooling water pumps (P_{COWP})
Frequency of cooling water pumps (F_{COWP})
On-off states of cooling water pumps (OOS_{COWP})
Number of working cooling water pumps (N_{COWP})
On-off states of chillers (OOS_{CH})
Power of chillers (P_{CH})
Load ratios of chillers (LR_{CH})
Number of working chillers (N_{CH})
Power of primary chilled water pumps (P_{PCHWP})
Frequency of primary chilled water pumps (F_{PCHWP})
On-off states of primary chilled water pumps (OOS_{PCHWP})
Number of working primary chilled water pumps (N_{PCHWP})
Chilled water valve on-off states of chillers (OOS_{CHWV_CH})
Chilled water valves of chillers
OOS_{CHWV_CH1}
OOS_{CHWV_CH2}
OOS_{CHWV_CH8}
Primary chilled water pumps
P_{PCHWP1}
P_{PCHWP2}
$P_{PCHWP10}$
F_{PCHWP1}
F_{PCHWP2}
$F_{PCHWP10}$
OOS_{PCHWP1}
OOS_{PCHWP2}
$OOS_{PCHWP10}$
N_{PCHWP}
Supply chilled water temperature of chillers (T_{SCHW_CH})
T_{SCHW_CH1}
T_{SCHW_CH2}
T_{SCHW_CH8}
Chilled water flow rate (FR_{CHW})
FR_{CHW}
Secondary chilled water pumps
N_{SCHWPA}
OOS_{SCHWP_A1}
F_{SCHWP_A1}
P_{SCHWP_A1}
OOS_{SCHWP_A2}
F_{SCHWP_A2}
P_{SCHWP_A2}
A zone
Temperature of total supply chilled water (T_{TSCHW})
T_{TSCHW}
TD_{CHW}
0.0 (By-pass pipe)
OOS_{SCHWP_D1}
F_{SCHWP_D1}
P_{SCHWP_D1}
D zone
N_{SCHWPD}
OOS_{SCHWP_D4}
F_{SCHWP_D4}
P_{SCHWP_D4}
Temperature of total return chilled water (T_{TRCHW})
T_{TRCHW}
Temperature difference of chilled water (TD_{CHW})
Number of working secondary chilled water pumps (N_{SCHWP})
Power of secondary chilled water pumps (P_{SCHWP})
Frequency of secondary chilled water pumps (F_{SCHWP})
On-off states of secondary chilled water pumps (OOS_{SCHWP})
0.0
0.5
1.0
A device node
A sensor node
A virtual variable node
A demand side node (regarded as a device node)

Fig. 10. Variable network graph of the chiller plant.

ler worked when the outdoor air temperature was low (between 10 °C and 15 °C). The temperature difference of chilled water was small (between 0 K and 3 K) in this situation, indicating that the cooling energy demand should be low in this situation. Therefore, the load ratio of #8 chiller was usually low or medium in this situation. Considering the efficiency of a chiller is relatively low

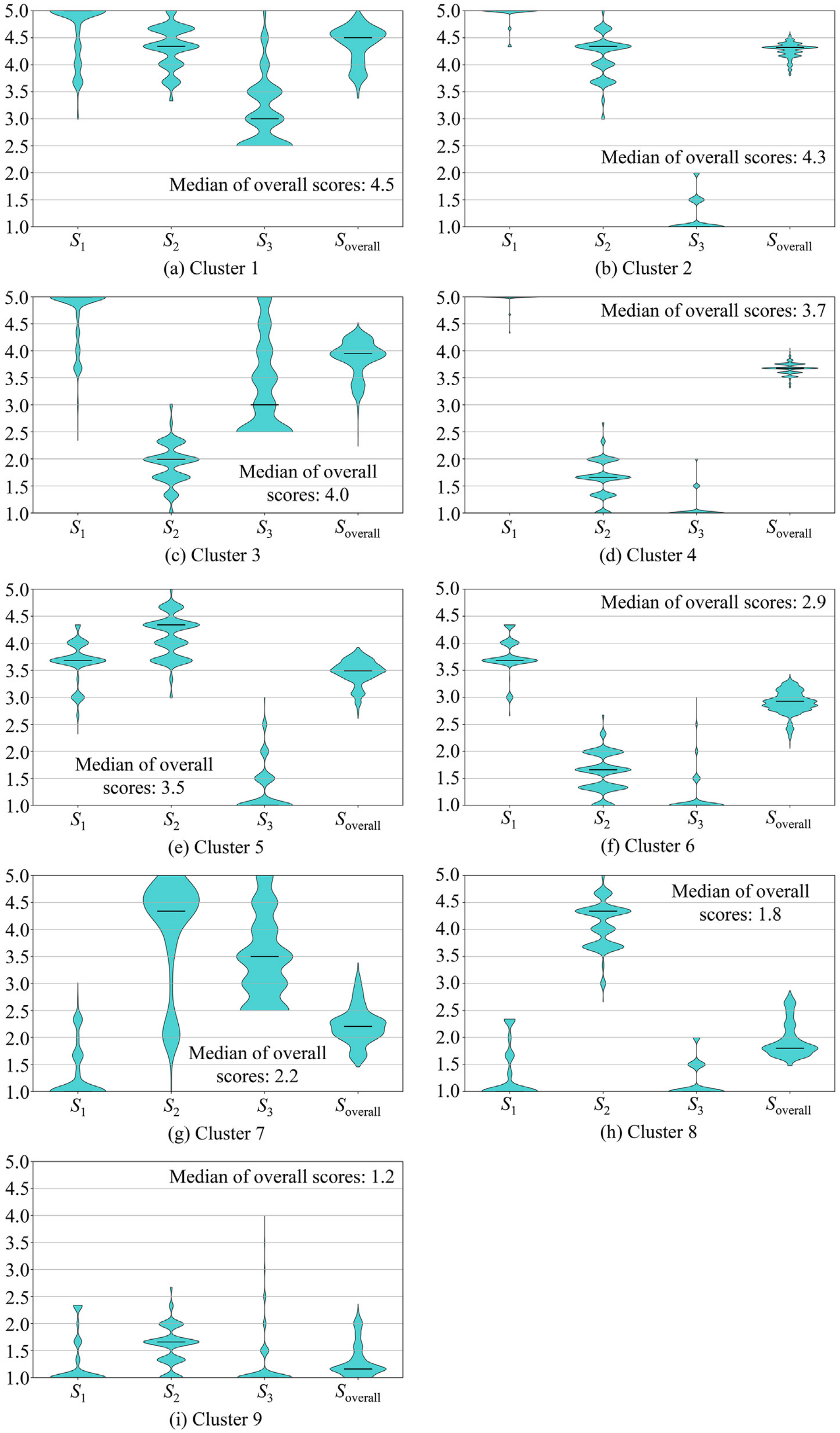

Fig. 11. Violin plots of the scores of criteria and the overall score in each cluster.

under part-load conditions, the operations of #8 chiller were energy-inefficient under low cooling energy demand.

Two solutions should be useful for energy efficiency enhancement of this chiller plant under low cooling energy demand. The first solution is adopting free cooling strategies under low cooling energy demand. As shown in Fig. 16, the temperature of total supply cooling water was relatively low when the outdoor air temperature was low. The cooling water with low temperature can be utilized as a free cooling source. Moreover, the cool outdoor air also can be utilized as a free cooling source. The second solution is using the intermittent operation strategies for chillers under low cooling energy demand.

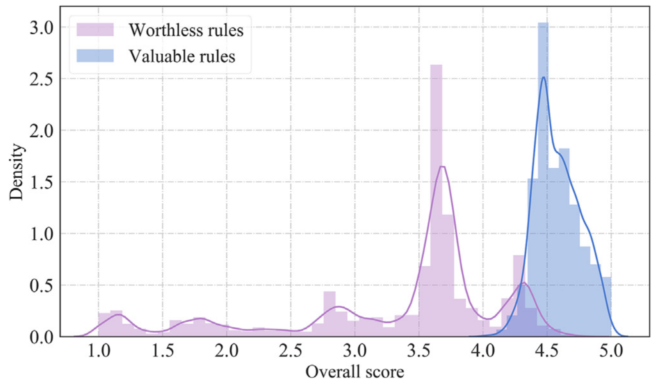

Fig. 12. Probability density distributions of overall scores of worthless and valuable association rules.

4. Practical application for operation anomaly detection of building energy systems

In this study, the operational data from an HVAC system' chiller plant have been analyzed successfully by the proposed method. The proposed method should also be applicable to other types of building energy systems, although these systems are different in terms of systems structures, control strategies, sensor configurations, and so on. The main reason is that the three criteria and six sub-criteria don't consider the difference in the system level. The criterion named stability attempts to remove the association rules related to transient or infrequent operation patterns that are worthless for each type of building energy systems. The criterion named physical correlation can remove the association rules related to unrelated or non-working devices. Such association rules are also worthless for other building energy systems. The criterion named statistical correlation aims to extract the association rules associated with variables with high statistical correlation. It is a common aspect considered in the previous studies. Hence, it also can be adopted in every type of building energy systems.

The proposed method improves the efficiency of rule extraction from massive amounts of operational data of building energy systems. The extracted rules can not only identify the operation anomalies occurred in the past, but also be further utilized to build expert rule bases for real-time operation anomaly detection. In general, it is hard for experts to enumerate all types of operation anomalies for various building energy systems, due to the systems' diversity and complexity. However, ARM can discover almost any types of operation patterns for each type of building energy systems. It can make the establishment of expert rule bases more convenient and more reliable.

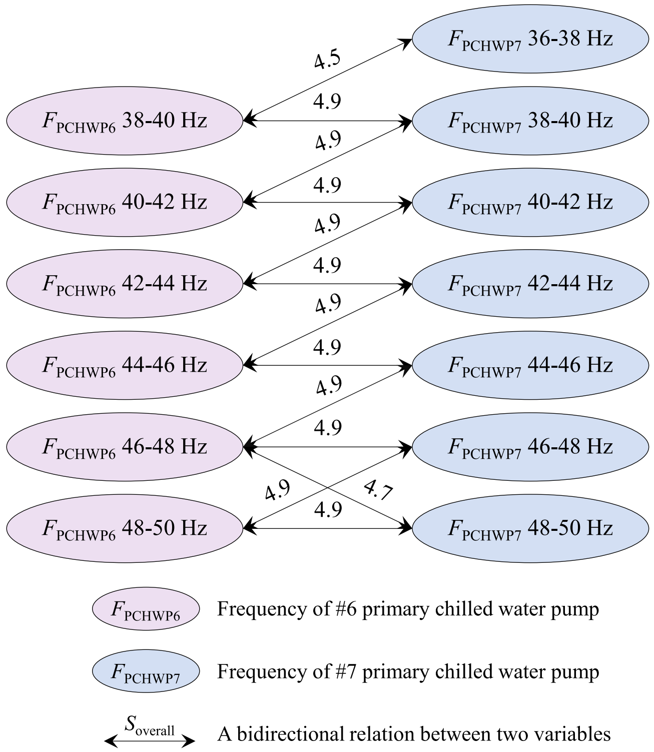

Fig. 14. Visualization of the bidirectional association rules related to the frequencies of #6 and #7 primary chilled water pumps.

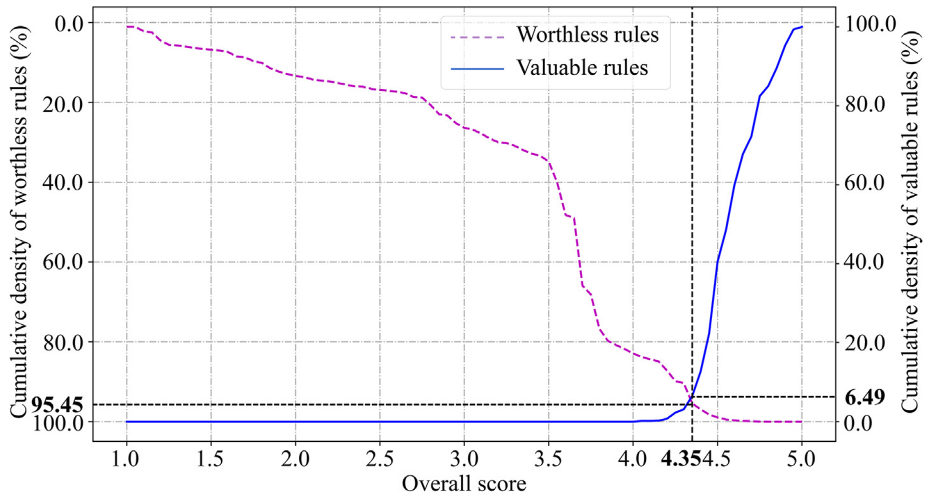

Fig. 13. Cumulative probability density distributions of overall scores of worthless and valuable association rules.

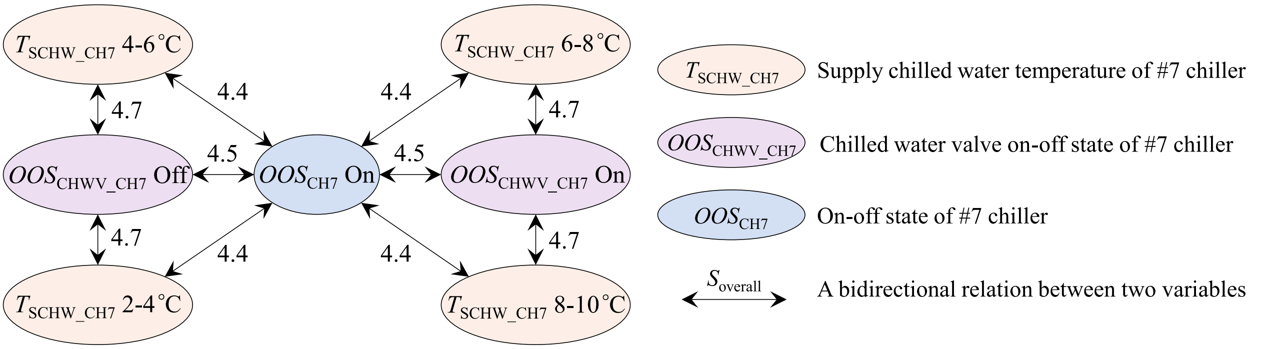
T_{SCHW_CH7} 4-6°C
T_{SCHW_CH7} 6-8°C
OOS_{CHWV_CH7} Off
OOS_{CH7} On
OOS_{CHWV_CH7} On
T_{SCHW_CH7} 2-4°C
T_{SCHW_CH7} 8-10°C
4.4
4.4
4.7
4.7
4.5
4.5
4.7
4.7
4.4
4.4
T_{SCHW_CH7} Supply chilled water temperature of #7 chiller
OOS_{CHWV_CH7} Chilled water valve on-off state of #7 chiller
OOS_{CH7} On-off state of #7 chiller
$S_{overall}$ A bidirectional relation between two variables

Fig. 15. Visualization of the bidirectional association rules related to #7 chiller.

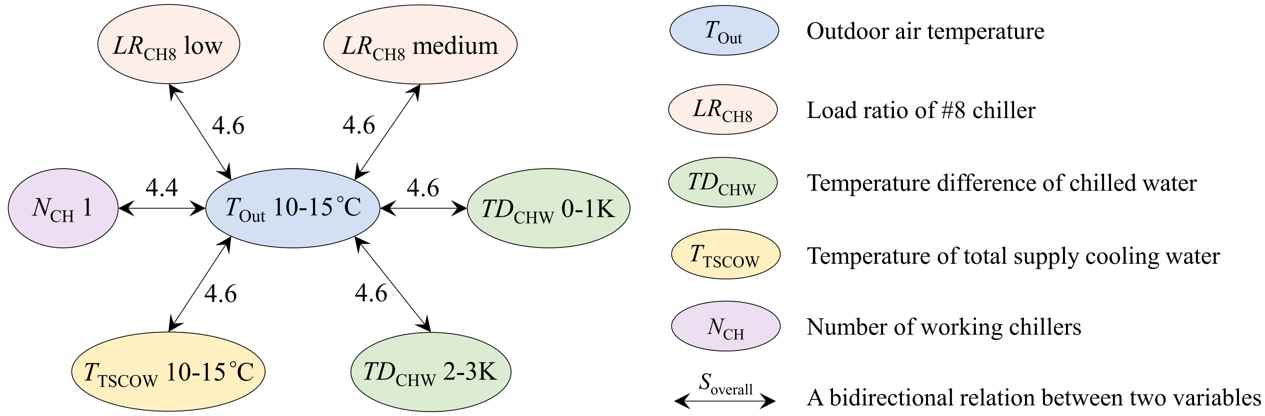
LR_{CH8} low
LR_{CH8} medium
4.6
4.6
N_{CH} 1
4.4
T_{Out} 10-15°C
4.6
TD_{CHW} 0-1K
4.6
4.6
T_{TSCOW} 10-15°C
TD_{CHW} 2-3K
T_{Out} Outdoor air temperature
LR_{CH8} Load ratio of #8 chiller
TD_{CHW} Temperature difference of chilled water
T_{TSCOW} Temperature of total supply cooling water
N_{CH} Number of working chillers
$S_{overall}$ A bidirectional relation between two variables

Fig. 16. Visualization of the bidirectional association rules related to the energy-inefficient operations of chillers under low cooling energy demand.

5. Conclusions

An AHP-based fuzzy post mining method is proposed in this study to grade the value of the discovered association rules for identifying the valuable association rules from the worthless ones. It grades the value of an association rule based on three criteria and six corresponding sub-criteria. The fuzzy set theory is adopted to grade each sub-criterion of an association rule for describing the uncertainties caused by imprecise judgments. AHP is then utilized to estimate the weight of each criterion/sub-criterion for getting the weighted overall score of an association rule. The higher the overall score of an association rule is, the more valuable the association rule is.

The AHP-based fuzzy post mining method is evaluated to grade the value of 117,636 association rules. To get ground truth, every association rule is labeled manually according to its worthiness for improving energy efficiency. For the worthless association rules, the proposed method can remove 95.45% of them successfully, and regard 4.55% of them as valuable ones wrongly. For the valuable association rules, the proposed method can identify 93.51% of them successfully, and regard 6.49% of them as worthless ones wrongly. Four existing indexes (*support*, *confidence*, *lift* and graph-based distance correlation) are further utilized as a traditional method for performance comparison with the proposed method. It is discovered that they can filter out 96.32% of the worthless association rules, but they identify 82.72% of the valuable association rules as worthless ones wrongly. It demonstrates that the performance of the proposed method is better than the traditional method.

The proposed post mining method has great potential to be applied in distributed energy systems, district heating and cooling systems, building lighting systems, and so on. Moreover, new criteria are also suggested to be developed in the future to further enhance the performance of the proposed method in isolating the valuable association rules from the worthless association rules. For instance, some specific working characteristics (such as specific control strategies) of a chiller plant can be selected as new criteria.

CRediT authorship contribution statement

Chaobo Zhang: Conceptualization, Methodology, Software, Data curation, Validation, Formal analysis, Visualization, Writing – original draft, Writing - review & editing. **Yang Zhao:** Supervision, Writing – original draft, Funding acquisition. **Jie Lu:** Investigation. **Tingting Li:** Investigation. **Xuejun Zhang:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the National Natural Science Foundation of China (No. 51978601) and the support of the National Key Research and Development Program of China (No. 2018YFE0116300).

References

[1] T. Hong, L. Yang, D. Hill, W. Feng, Data and analytics to inform energy retrofit of high performance buildings, Appl. Energy 126 (2014) 90–106, https://doi.org/10.1016/j.apenergy.2014.03.052.

[2] R.Z. Homod, H. Togun, H.J. Abd, K.S.M. Sahari, A novel hybrid modelling structure fabricated by using Takagi-Sugeno fuzzy to forecast HVAC systems energy demand in real-time for Basra city, Sustain. Cities Soc. 56 (2020) 102091, https://doi.org/10.1016/j.scs.2020.102091.

[3] V. Vakiloroaya, B. Samali, A. Fakhar, K. Pishghadam, A review of different strategies for HVAC energy saving, Energy Convers. Manage. 77 (2014) 738–754, https://doi.org/10.1016/j.enconman.2013.10.023.

[4] D. Lee, C.-C. Cheng, Energy savings by energy management systems: a review, Renew. Sustain. Energy Rev. 56 (2016) 760–777, https://doi.org/10.1016/j.rser.2015.11.067.

[5] Z. Ma, R. Yan, K. Li, N. Nord, Building energy performance assessment using volatility change based symbolic transformation and hierarchical clustering, Energy Build. 166 (2018) 284–295, https://doi.org/10.1016/j.enbuild.2018.02.015.

[6] M. Kordestani, A.A. Safavi, M. Saif, Recent survey of large-scale systems: architectures, controller strategies, and industrial applications, IEEE Syst. J. (2021) 1–14, https://doi.org/10.1109/JSYST.2020.3048951.

[7] T. Daixin, X. Hongwei, Y. Huijuan, Y. Hao, H. Wen, Optimization of group control strategy and analysis of energy saving in refrigeration plant, Energy Built Environ. (2021), https://doi.org/10.1016/j.enbenv.2021.05.006.

[8] I. Ganchev, A. Taneva, K. Kutryanski, M. Petrov, Decoupling fuzzy-neural temperature and humidity control in HVAC systems, IFAC-PapersOnLine 52 (2019) 299–304, https://doi.org/10.1016/j.ifacol.2019.12.539.

[9] M. Kordestani, M. Saif, M.E. Orchard, R. Razavi-Far, K. Khorasani, Failure prognosis and applications—a survey of recent literature, IEEE Trans. Reliab. 70 (2) (2021) 728–748, https://doi.org/10.1109/TR.2410.1109/TR.2019.2930195.

[10] M. Mousavi, M. Moradi, A. Chaibakhsh, M. Kordestani, M. Saif, Ensemble-based fault detection and isolation of an industrial gas turbine, in: Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020, pp. 2351–2358. doi:10.1109/SMC42975.2020.9282904.

[11] M. Kordestani, M. Rezamand, M. Orchard, R. Carriveau, D.S.K. Ting, M. Saif, Planetary gear faults detection in wind turbine gearbox based on a ten years historical data from three wind farms, IFAC-PapersOnLine 53 (2020) 10318–10323, https://doi.org/10.1016/j.ifacol.2020.12.2767.

[12] Y. Zhao, T. Li, X. Zhang, C. Zhang, Artificial intelligence-based fault detection and diagnosis methods for building energy systems: advantages, challenges and the future, Renew. Sustain. Energy Rev. 109 (2019) 85–101, https://doi.org/10.1016/j.rser.2019.04.021.

[13] Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, J. Li, A review of data mining technologies in building energy systems: load prediction, pattern identification, fault detection and diagnosis, Energy Built Environ. 1 (2) (2020) 149–164, https://doi.org/10.1016/j.enbenv.2019.11.003.

[14] C. Miller, Z. Nagy, A. Schlueter, A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings, Renew. Sustain. Energy Rev. 81 (2018) 1365–1377, https://doi.org/10.1016/j.rser.2017.05.124.

[15] M.S. Mirnaghi, F. Haghighat, Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: a comprehensive review, Energy Build. 229 (2020) 110492, https://doi.org/10.1016/j.enbuild.2020.110492.

[16] C. Zhang, Y. Zhao, Y. Zhou, X. Zhang, T. Li, A real-time abnormal operation pattern detection method for building energy systems based on association rule bases, Build. Simul. (2021), https://doi.org/10.1007/s12273-021-0791-x.

[17] Z. Yu, F. Haghighat, B.C.M. Fung, L. Zhou, A novel methodology for knowledge discovery through mining associations between building operational data, Energy Build. 47 (2012) 430–440, https://doi.org/10.1016/j.enbuild.2011.12.018.

[18] Z. Yu, B.C.M. Fung, F. Haghighat, Extracting knowledge from building-related data—a data mining framework, Build. Simul. 6 (2) (2013) 207–222, https://doi.org/10.1007/s12273-013-0117-8.

[19] F. Xiao, C. Fan, Data mining in building automation system for improving building operational performance, Energy Build. 75 (2014) 109–118, https://doi.org/10.1016/j.enbuild.2014.02.005.

[20] C. Fan, F. Xiao, C. Yan, A framework for knowledge discovery in massive building automation data and its application in building diagnostics, Autom. Constr. 50 (2015) 81–90, https://doi.org/10.1016/j.autcon.2014.12.006.

[21] C. Fan, F. Xiao, Mining big building operational data for improving building energy efficiency: a case study, Build. Serv. Eng. Res. Technol. 39 (1) (2018) 117–128, https://doi.org/10.1177/0143624417704977.

[22] G. Li, Y. Hu, H. Chen, H. Li, M. Hu, Y. Guo, J. Liu, S. Sun, M. Sun, Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions, Appl. Energy 185 (2017) 846–861, https://doi.org/10.1016/j.apenergy.2016.10.091.

[23] C. Zhang, X. Xue, Y. Zhao, X. Zhang, T. Li, An improved association rule mining-based method for revealing operational problems of building heating, ventilation and air conditioning (HVAC) systems, Appl. Energy 253 (2019) 113492, https://doi.org/10.1016/j.apenergy.2019.113492.

[24] S. Qiu, F. Feng, Z. Li, G. Yang, P. Xu, Z. Li, Data mining based framework to identify rule based operation strategies for buildings with power metering system, Build. Simul. 12 (2) (2019) 195–205, https://doi.org/10.1007/s12273-018-0472-6.

[25] P. Xue, Z. Zhou, X. Fang, X. Chen, L. Liu, Y. Liu, J. Liu, Fault detection and operation optimization in district heating substations based on data mining techniques, Appl. Energy 205 (2017) 926–940, https://doi.org/10.1016/j.apenergy.2017.08.035.

[26] X. Zhou, W. Lin, P. Cui, Z. Ma, T. Huang, An unsupervised data mining strategy for performance evaluation of ground source heat pump systems, Sustain. Energy Technol. Assess. 46 (2021) 101255, https://doi.org/10.1016/j.seta.2021.101255.

[27] H.B. Gunay, W. Shen, C. Yang, Text-mining building maintenance work orders for component fault frequency, Building Res. Inform. 47 (5) (2019) 518–533, https://doi.org/10.1080/09613218.2018.1459004.

[28] S. Dutta, H.B. Gunay, S. Bucking, Benchmarking operational performance of buildings by text mining tenant surveys, Sci. Technol. Built Environ. 27 (6) (2021) 741–755, https://doi.org/10.1080/23744731.2020.1851545.

[29] C. Fan, F. Xiao, H. Madsen, D. Wang, Temporal knowledge discovery in big BAS data for building energy management, Energy Build. 109 (2015) 75–89, https://doi.org/10.1016/j.enbuild.2015.09.060.

[30] M.S. Piscitelli, D.M. Mazzarelli, A. Capozzoli, Enhancing operational performance of AHUs through an advanced fault detection and diagnosis process based on temporal association and decision rules, Energy Build. 226 (2020) 110369, https://doi.org/10.1016/j.enbuild.2020.110369.

[31] C. Fan, Y. Sun, K. Shan, F. Xiao, J. Wang, Discovering gradual patterns in building operations for improving building energy efficiency, Appl. Energy 224 (2018) 116–123, https://doi.org/10.1016/j.apenergy.2018.04.118.

[32] C. Fan, F. Xiao, Z. Li, J. Wang, Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: a review, Energy Build. 159 (2018) 296–308, https://doi.org/10.1016/j.enbuild.2017.11.008.

[33] Z. Yu, F. Haghighat, B.C.M. Fung, E. Morofsky, H. Yoshino, A methodology for identifying and improving occupant behavior in residential buildings, Energy 36 (11) (2011) 6596–6608, https://doi.org/10.1016/j.energy.2011.09.002.

[34] Y. Zhao, C. Zhang, L. Cao, Post-mining of association rules: Techniques for effective knowledge extraction, IGI Global, Hershey, 2009, https://10.4018/978-1-60566-404-0.

[35] C. Zhang, Y. Zhao, T. Li, X. Zhang, J. Luo, A comprehensive investigation of knowledge discovered from historical operational data of a typical building energy system, J. Building Eng. 42 (2021) 102502, https://doi.org/10.1016/j.jobe.2021.102502.

[36] C. Zhang, Y. Zhao, T. Li, X. Zhang, A post mining method for extracting value from massive amounts of building operational data, Energy Build. 223 (2020), https://doi.org/10.1016/j.enbuild.2020.110096 110096.

[37] T.L. Saaty, A scaling method for priorities in hierarchical structures, J. Math. Psychol. 15 (3) (1977) 234–281, https://doi.org/10.1016/0022-2496(77)90033-5.

[38] M. Shahrestani, R. Yao, G.K. Cook, A fuzzy multiple attribute decision making tool for HVAC&R systems selection with considering the future probabilistic climate changes and electricity decarbonisation plans in the UK, Energy Build. 159 (2018) 398–418, https://doi.org/10.1016/j.enbuild.2017.10.089.

[39] E.W. Dijkstra, A note on two problems in connexion with graphs, Numer. Math. 1 (1) (1959) 269–271, https://doi.org/10.1007/BF01386390.

[40] R. Bellman, On a routing problem, Q. Appl. Math. 16 (1) (1958) 87–90, https://doi.org/10.1090/qam/1958-16-0110.1090/qam/102435.

[41] R.W. Floyd, Algorithm 97: shortest path, Commun ACM 5 (1962) 345, https://doi.org/10.1145/367766.368168.

[42] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72–101, https://doi.org/10.2307/1412159.

[43] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, Detecting novel associations in large data sets, Science 334 (6062) (2011) 1518–1524, https://doi.org/10.1126/science:1205438.

[44] G. Sun, J. Li, J. Dai, Z. Song, F. Lang, Feature selection for IoT based on maximal information coefficient, Future Generation Comput. Syst. 89 (2018) 606–616, https://doi.org/10.1016/j.future.2018.05.060.

[45] T.L. Saaty, How to make a decision: the analytic hierarchy process, INFORMS J. Appl. Anal. 24 (6) (1994) 19–43, https://doi.org/10.1287/inte.24.6.19.

[46] F. Omar, S.T. Bushby, R.D. Williams, Assessing the performance of residential energy management control algorithms: multi-criteria decision making using the analytical hierarchy process, Energy Build. 199 (2019) 537–546, https://doi.org/10.1016/j.enbuild.2019.07.033.

[47] S.M. Sajjadian, M. Jafari, D. Pekaslan, An expandable, contextualized and data-driven indoor thermal comfort model, Energy Built Environ. 1 (4) (2020) 385–392, https://doi.org/10.1016/j.enbenv.2020.04.005.

[48] H. Chaouch, C. Çeken, S. Arı, Energy management of HVAC systems in smart buildings by using fuzzy logic and M2M communication, J. Building Eng. 44 (2021) 102606, https://doi.org/10.1016/j.jobe.2021.102606.

[49] W. Pedrycz, M. Song, Analytic hierarchy process (AHP) in group decision making and its optimization with an allocation of information granularity, IEEE Trans. Fuzzy Syst. 19 (3) (2011) 527–539, https://doi.org/10.1109/TFUZZ.2011.2116029.

[50] S. Hougardy, The Floyd-Warshall algorithm on graphs with negative cycles, Inform. Process. Lett. 110 (8-9) (2010) 279–281, https://doi.org/10.1016/j.ipl.2010.02.001.

[51] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. 3 (1) (1974) 1–27, https://doi.org/10.1080/03610927408827101.

[52] E.G. Hansen, Parallel operation of variable speed pumps in chilled water systems, ASHRAE J. 37 (1995) 34–38.