Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 111 (2017) 1070 – 1078

Energy Procedia

8th International Conference on Sustainability in Energy and Buildings, SEB-16, 11-13 September 2016, Turin, ITALY

Assessment of building operational performance using data mining techniques: a case study

Cheng Fan, Fu Xiao*

Department of Building Services Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Abstract

Today's buildings are not only energy intensive, but also information intensive. Massive amounts of operational data are available for knowledge discovery. Data mining (DM) has excellent ability in extracting insights from massive data. This paper performs a case study on the assessment of building operational performance using DM techniques. Typical DM techniques are compared and considerations for choosing specific DM techniques for the case study are presented. The methodology developed has been applied to analyze the data retrieved from a university building in Hong Kong. Useful insights have been obtained to identify typical operation patterns and energy conservation opportunities.

Peer-review under responsibility of KES International.

Keywords: Building Energy Conservation; Building Automation System; Data Mining; Big Data; Intelligent Building.

1. Introduction

Buildings have become one of the largest energy consumers around the world. The energy saving potential in building operations is huge due to the widespread occurrence of equipment degradation, faults in system components, and deficiencies in control strategies in buildings. Advanced technologies, such as the building automation system (BAS), have been integrated with modern buildings to facilitate the real-time monitoring and controls over building operations. Massive amounts of building operational data are collected and stored in BAS, from which valuable insights can be extracted to enhance the building operational performance. Nevertheless, building data are far from being fully utilized, mainly due to the lack of methods and tools for handling those big

* Corresponding author. Tel.: 852-27664194; fax: 852-27657198
E-mail address: linda.xiao@polyu.edu.hk

Peer-review under responsibility of KES International.
doi:10.1016/j.egypro.2017.03.270

data. Conventional methods for using the building data, which primarily rely on physical principles, statistics and engineering expertise, are neither efficient nor effective in discovering potentially useful yet previously unknown knowledge from massive BAS data sets. Advanced methodologies and tools are urgently needed in the building field to tackle the big data challenge.

Data mining (DM) technology is a promising solution and renowned for its excellent ability in extracting useful insights from massive data sets. It has been widely used in various industries, including the financial services, retails, health care, and even counter-terrorism [1, 2]. In general, DM techniques can be classified into two groups, i.e., supervised and unsupervised learning. Supervised learning techniques aims to perform regression or classification based on the relationships discovered between input and output variables. The knowledge discovered is usually represented using various models. Supervised learning techniques have been widely applied for energy consumption prediction [3-6] and fault detection and diagnosis [7-10] in the building field. One intrinsic limitation of supervised learning is that it needs reliable training data, which are very hard to obtain in building operations, particularly data under fault conditions. By contrast, unsupervised learning doesn't have such a need and it focuses on discovering the intrinsic structures, correlations and associations in the data. Moreover, it requires less domain expertise which makes it more preferable in real applications to discover new knowledge. The knowledge obtained using unsupervised DM techniques is usually in the form of data clusters, association rules, or anomalies.

This paper performs a case study on extracting useful knowledge from massive building operational data using DM techniques and their potential applications in building energy management. The methodology is derived from the generic data analytic framework, which was proposed in our previous study [11]. The main DM techniques adopted are decision trees and association rule mining. The methodology has been applied to analyze the data retrieved from one building in the Hong Kong Polytechnic University. The results show that useful insights can be obtained for enhancing building energy efficiency.

2. Research Methodology

2.1. Research outline

The knowledge gap between building professionals and advanced analytics motivated us to develop a generic DM-based analytic framework for analyzing big building operational data. Based on extensive investigation of popular DM techniques and deep understanding of building operations, a framework has been proposed in our previous paper [11]. The framework contains 4 phases, i.e., data exploration, data partitioning, knowledge discovery and post-mining. The data exploration phase mainly aims to enhance the data quality and prepares the data into suitable formats for the following data analysis. The data partitioning phase intends to improve the reliability and sensitivity of the knowledge discovered by dividing the building operational data into several groups according to the characteristics of building operations. Various DM techniques can be adopted to extract knowledge at the knowledge discovery phase. Domain expertise is involved in the post-mining phase to interpret, select and apply potentially useful knowledge.

The methodology adopted in this paper is derived from the framework. The clustering analysis method and the decision tree method are compared and the latter is chosen for data partitioning. The quantitative association rule mining is applied for knowledge discovery. The details are introduced in the following sub-sections.

2.2. Data partitioning

Building operations are highly complicated due to the constantly changing indoor and outdoor conditions. It is therefore not wise to treat the building operational data as a whole for data analysis, as it will negatively affect the reliability and sensitivity of knowledge discovered. Typical building operational data are stored in a two-dimensional data table, in which each column represents a variable and each row stores the values of different variables sampled at the same instant of time. Data partitioning refers to the process of dividing the entire data table into several subsets, each containing a number of rows.

Two types of methods are suitable for partitioning building operational data. The first is to treat each row as an observation and then grouping observations based on their similarities which can be evaluated by Euclidean distance

or Cosine distance. The clustering analysis is the effective method to perform this task. It divides the data into several clusters with the aim of minimizing the between-cluster similarities and maximizing the within-cluster similarities. The variables in building operational data are usually of different scales, e.g., the power consumption of a chiller may range from 0~1000 kW while the supplied chilled water temperature may range from 5 ~ 10^0C. Therefore, one essential for this method is to normalize the data before calculating their similarities. It is worth mentioning that due to the curse of dimensionality, the similarity measures may become meaningless when the variable number is large [12]. Therefore, users may have to select a subset of variables as inputs for clustering analysis when the variable number is large. This subset should be able to reflect the changes in building operations. The main drawback of this method is that the result is more like a black-box model. The results lack interpretability, as the only output is the clustering membership of each observation. Further data exploration has to be carried out if users want to know the data characteristics in each cluster.

The other method is to partition the massive data sets according to one typical variable which can represent the building operation characteristics and is also a major concern, e.g., the building power consumption. The decision tree method can fulfill this task and the results obtained are highly interpretable. An example of decision tree model is shown in Fig. 1. The model depicts the relationship between building cooling load, outdoor temperature and indoor occupancy ratio. Nodes 3 to 5 are called terminal nodes, which present the prediction result of cooling load under different scenarios. For instance, Node 3 indicates that the building cooling load is *Low* if the outdoor temperature is no more than 24^{o}C and the indoor occupancy ratio is no more than 0.5. The decision tree model is highly interpretable and it offers clues on how to partition the data. In this study, the decision method is applied for the data partitioning task.

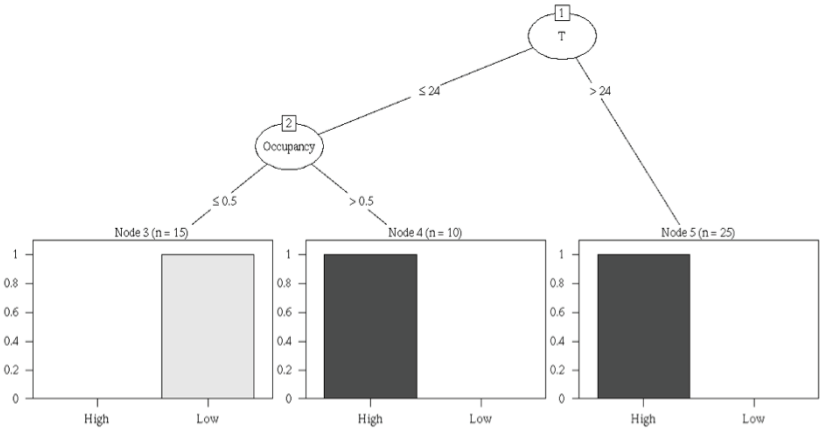

Fig. 1. An example decision tree model

2.3. Knowledge discovery methods

Investigating the relationships between different variables is the main approach for knowledge discovery. Association rule mining is popular method to mine associations. An association rule $A \rightarrow B$ states that if A happens, then B happens, where A is the antecedent and B is the consequence. Two thresholds are required to perform the association rule mining, i.e., support and confidence. The support of $A \rightarrow B$ is the joint probability of A and B while the confidence of $A \rightarrow B$ is the conditional probability of B given A. The number of association rules obtained decreases with the increase in the support and confidence thresholds. One drawback of association rule mining is that the number of association rules obtained is too large which results in heavy load in manual inspection of useful rules. In such a case, users may use some statistics to measure the rule interestingness. For instance, the lift value is defined as the ratio between the rule confidence and the support of the consequence [2]. If

the lift value is larger than 1, it indicates the presence of consequence is positively affected by the presence of antecedence and vice versa. A lift value of 1 indicates that the antecedent and consequent are independent from each other and therefore, the rule is of little value.

In general, association rule mining algorithms only works with categorical variables. The majority of building operational data is numeric and therefore, discretization becomes necessary. Data discretization for building operational data is a challenging problem, as variables usually have their own behavior and the optimal discretization levels are hard to define without prior knowledge. For instance, the power consumption data of a two-level speed fan exhibits a typical bimodal distribution. It is evident that they can be discretized into three categories indicating the operating conditions of *Idle*, *Low* and *High*. By contrast, the flow rates of a variable speed pump may have a more uniform distribution, and it is difficult to define the number of categories and the breakpoints. Such discretization usually leads to information loss and may severely downgrades the mining performance.

To avoid the drawbacks of discretization, some algorithms, so called quantitative association rule mining, have been proposed so that the association rule mining can be performed on both numeric and categorical variables. . This study adopts the QuantMiner [13] as the mining algorithm. If the variable is numeric, then an interval is automatically identified considering the rule gain and the coverage of the interval identified. The identified interval is then used to create categorical values. The rule gain is calculated using Eq. 1, where *MinConf* refers to the minimal confidence threshold. Genetic algorithm is applied to identify the interval by maximizing a fitness function, as shown in Eq. 2, where A_{num} is the number of numeric variables in the rule; I_{Ai} is the interval of A_i; $size(A_i)$ is the range of A_i; $size(I_{Ai})$ is the length of the identified interval. The algorithm prefers to select rules with large gains and small intervals.

$$Gain(A \rightarrow B) = Support(A,B) - MinConf \times Support(A) \quad (1)$$

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

3. Case study

3.1. Description of building, systems and building operational data

The Phase 5 building located in the campus of the Hong Kong Polytechnic University is selected for case study. It mainly consists of offices, classrooms and a computer data center. The gross floor area is approximately 11,000m^2, of which about 8,500m^2 are air-conditioned spaces.

The data collected for Phase 5 come from two sources. One set comes from the power meter, which measures the total building electricity consumption at the interval of 30-minute. The other set comes from the Building Automation System (BAS), which monitors the performance of the chiller plant of Heating, Ventilation and Air-Conditioning (HVAC) system at the interval of 1-minute. The chiller plant contains 2 water-cooled chillers (denoted as CH-1 and CH-2) and 2 cooling towers (denoted as CT-1 and CT-2). Chillers are connected in parallel and the chilled water is distributed using 3 variable speed driver (VSD) pumps (denoted as PCHWP-1 to 3). The condenser water is circulated between chillers and 2 cooling towers using 3 VSD pumps (denoted as CDWP-1 to 3). The design specifications of main chiller plant components are summarized in Table 1.

One-year data retrieved from the BAS (from January 2014 to January 2015) are used for analysis. The data have 52 variables, including the building electricity consumption and measurements of the water-side HVAC system, e.g., temperature, flow rate and pressure.

Table 1. Specification of the chiller plant

Components	Number	Remarks	Power (*kW*)
CH-1/CH-2	2	Cooling capacity: 1050 kW	338.3
PCHWP-1~3	3	VSD, Flow: 50.4 *l/s*	26.7
CDWP-1~3	3	VSD, Flow: 62.8 *l/s*	18.5
CT-1/2	2	VSD	11

3.2. Identifying typical building operation patterns

Building power consumption, which is sensitive to the outdoor and indoor conditions, is a typical variable related to the building operations. It can be used as an indicator of different building operation patterns. As introduced in section 2.2, the decision tree method is adopted in data partitioning. In this study, the building power consumption is considered as the output variable and the time variables, such as the *Year*, *Month*, *Day*, *Hour*, *Minute* and *Day type* as the input variables. The indoor variables, such as the occupant number, are not used as inputs because, firstly, those data are not available due to the lack of measurement instruments; secondly, they are not necessary considering that the time and day type determines how people use the spaces.

A decision tree model is constructed using the electricity consumption data in the whole year, which is shown in Fig. 2. The model selects the *Month*, *Hour* and *Day type* as the splitting variable. Starting from Node 1, the model first picks the *Hour* as the splitting variable and the splitting criterion is {0, 1, 2, 3, 4, 5, 6, 7, 22, 23} and {8 to 21}. The result matches our domain knowledge as it corresponds to the non-peak and peak hours. The lectures normally start at 8:30am and end at 9:30pm. Node 2 divided the data based on the *Month*, one is {1, 2, 3, 12} and the other is {4, 5, 6, 7, 8, 9, 10, 11}. The first set corresponds to the cool and less humid seasons while the second refers to the hot and more humid seasons in Hong Kong. Node 4 selects the *Day type* and the partitioning is basically made based on weekdays and weekends.

The decision model constructed provides evident clues on data partitioning. Rather than dividing the whole data into 4 data groups according to the terminal nodes, the splitting criteria generated at Nodes 1, 2 and 5 are used. As a result, the entire data sets are partitioned into 8 groups, as shown in Table 2. Fig. 3 presents the boxplots of building electricity consumption in each partition. It is apparent that the building power consumption in each group presents different distribution, especially when it belongs to the peak hours. The power consumptions during non-peak hours on weekdays and weekends in the same cool or hot season are generally the same, e.g. Group 1 and Group 3, Group 5 and Group 7 as shown in Fig. 1. It is worth mentioning that the data in Group 4, 5 and 7 are quite similar, which states that the power consumption during peak hours on weekends in cool seasons is similar to that during non-peak hours in hot seasons. Conventional approach may easily classify those observations into one group. The decision tree method performs better in this case which can improve the sensitivity and reliability of the knowledge discovered.

Table 2. Data partition details

Partitions	Month	Day type	Hour
1	{1,2,3,12}	{Monday to Friday}	{0,1,2,3,4,5,6,7,22,23}
2	{1,2,3,12}	{Monday to Friday}	{8 to 21}
3	{1,2,3,12}	{Saturday, Sunday}	{0,1,2,3,4,5,6,7,22,23}
4	{1,2,3,12}	{Saturday, Sunday}	{8 to 21}
5	{4 to 11}	{Monday to Friday}	{0,1,2,3,4,5,6,7,22,23}
6	{4 to 11}	{Monday to Friday}	{8 to 21}
7	{4 to 11}	{Saturday, Sunday}	{0,1,2,3,4,5,6,7,22,23}
8	{4 to 11}	{Saturday, Sunday}	{8 to 21}

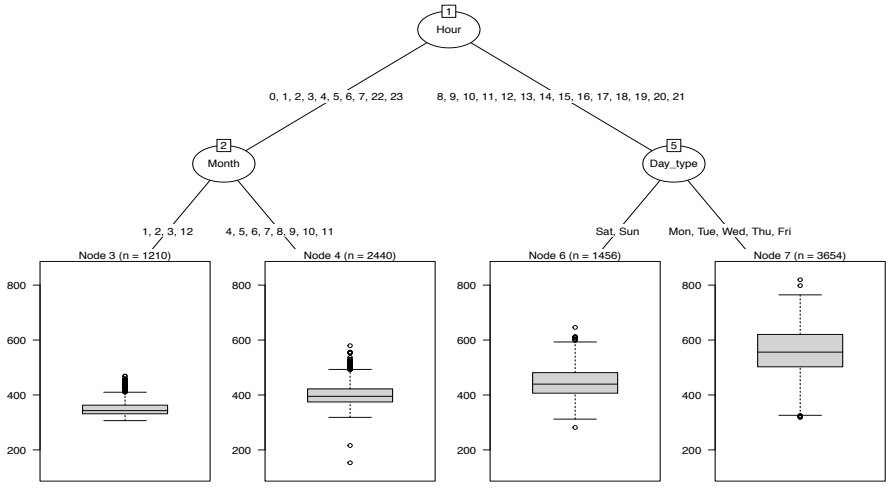

Fig. 2. Decision tree model for building electricity consumption

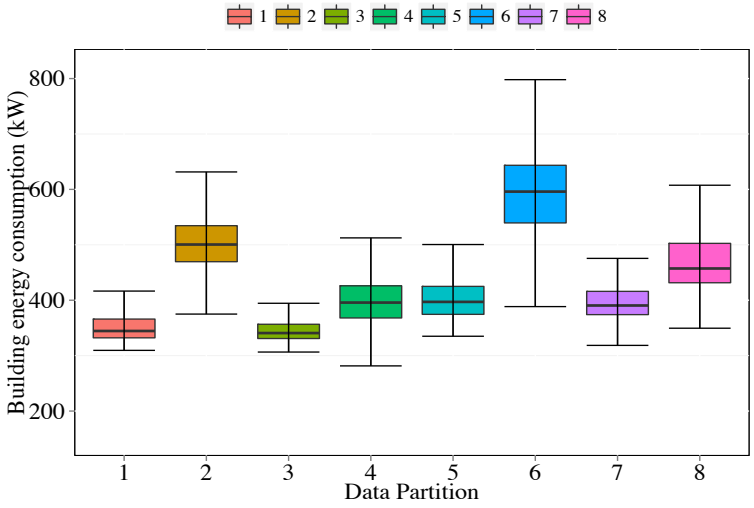

Fig. 3. Boxplots of building electricity consumption in each partition

3.3. Discovering associations in building operational data

As introduced in section 2.3, the QuantMiner algorithm is adopted to discover the associations in each group of data separately. For the convenience of interpreting the rules obtained, both sides of the rule, i.e. the antecedent and the consequence, are constrained to have one variable only. The parameters for the genetic algorithm are set as follows: 250 as the population size, 100 as iteration number, 50% as crossover rate and 40% as mutation rate. These parameters are set according to the suggestions of [13]. The support and confidence thresholds are set as 15% and 90% respectively. In general, the confidence threshold should be set no less than 0.8 to ensure the quality of association rules. The support values can be set according to the user's actual need. A small support threshold helps

to discover less frequent associations. It can be used to discover atypical associations in building operations. However, a smaller support threshold may lead to a dramatic increase in the association rules obtained, which makes the post-mining phase more time-consuming.

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

Table 3. Example quantitative association rules considering data Group 5 and 6

No.	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Data partition
1	CH_No = 0	CDW_Flow in [0.0, 1.0]	87.9	98.8	1.21	5
2	CT2_Status = Off	CT2_MotorSpeed in [0.0, 0.6]	89.6	100	1.12	5
3	CH2_status = Off	CH2_CHW_Flow in [-0.1, 0.3]	73.4	99.5	1.36	6

4. Applications

4.1. Chilled water and condensing water distribution system

Two examples rules presented in Table 4 indicate that when one chiller is switched on, its chilled water and condensing water flow rates become nearly constant. By checking the actual motor speed of PCHWP and CDWP, it is found out that the motor frequency was maintained at 40Hz during operation, which means the variable speed pumps don't operate at variable speed. The insights obtained helps to spot the energy conservations in actual operations, as control strategy should be developed to optimize the pressure set-point for pump speed control according to the actual cooling load and weather conditions.

Table 4. Associations in chilled water and condensing water flow rates

No.	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Data Group
1	CH1_Status = On	CH1_CHW_Flow in [47.1, 51.3]	60.1	99.5	1.65	6
2	CH2_Status = On	CH2_CDW_Flow in [46.8, 51.9]	26.1	99.7	3.80	6

4.2. Chiller control strategy

The rules in Table 5 depict the supplied chilled water temperature when one chiller is switched on. The intervals identified for the chilled water supply temperature are quite narrow. By checking with the building operation staff, it turns out that the set-point was set fixed as 7^{o}C. Considering that the chilled water supply temperature has a huge impact on the chiller power consumption [14], it is suggested to develop a temperature reset scheme to regulate the chilled water supply temperature.

Table 5. Associations in chiller operation

No.	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Data group
1	CH1_Status = On	CH1_CHW_ST in [6.8, 7.3]	60.0	99.4	1.66	6
2	CH2_Status = On	CH2_CHW_ST in [6.8, 7.8]	26.0	99.2	3.81	6

4.3. Cooling tower control strategy

The rules in Table 6 indicate that the cooling tower fan speed was maintained at around 35Hz during operations. In this case, an optimal condenser inlet water temperature set-point reset scheme should be developed. It should be able to provide fan speed set-points according to the ambient and working conditions to minimize the overall energy use of chillers and cooling tower fans.

Table 6. Associations in cooling tower operation

No.	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Data partition
1	CT1_Status = On	CT1_MotorFrequency in [35.0, 35.6]	84.7	99.5	1.17	6
2	CT2_Status = On	CT2_MotorFrequency in [35.2, 35.7]	82.5	99.6	3.21	6

5. Conclusion

This paper presents a case study on the effective utilization of massive building operational data through DM techniques. The methodology is derived from the generic framework proposed in our previous work. It contains three main phases, i.e., data partitioning, knowledge discovery and post-mining. Decision tree method is applied to provide clues on data partitioning. One unsupervised DM technique, i.e., association rule mining, is adopted as the main tool in the knowledge discovery process. The main benefits of using association rule mining include requiring no training data and little prior knowledge, having the ability to discover previous unknown knowledge, and the results are highly interpretable. Considering that the majority of building operational data is numeric and data discretization can be troublesome and time-consuming, a quantitative association rule mining algorithm, QuantMiner, is applied. Domain expertise is involved in the post-mining phase for knowledge interpretation, selection and application. Valuable knowledge has been extracted from a data set retrieved from a university building in Hong Kong. Frequent operation patterns of HVAC systems have been discovered revealing energy conservation opportunities in pumps, chillers and cooling tower operations.

Acknowledgements

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

References

[1] Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications-A decade review from 2000 to 2011. Expert Syst Appl 2012;39:11303-11.
[2] Witten IH, Frank E, Hall MA. Data mining: Practical machine learning tools and techniques. 3rd ed. Massachusetts: Morgan Kaufmann; 2011.
[3] Fan C, Xiao F. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. Appl Energ 2014;127:1-10.
[4] Zhao DY, Zhong M, Zhang X, Su X. Energy consumption predicting model of VRV (variable refrigerant volume) system in office buildings based on data mining. Energy 2016;102:660-8.
[5] Le Cam M, Daoud A, Zmeureanu R. Forecasting electric demand of supply fan using data mining techniques. Energy 2016;101:541-57.
[6] Zeng YH, Zhang ZJ, Kusiak A. Predictive modeling and optimization of a multi-zone HVAC system with data mining and firefly algorithms. Energy 2015;86:393-402.
[7] Capozzoli A, Lauro F, Khan I. Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Syst Appl 2015;42:4324-38.
[8] Fan C, Xiao F, Madsen H, Wang D. Temporal knowledge discovery in big BAS data for building energy management. Energ Buildings 2015;109:75-89.
[9] Xiao F, Fan C. Data mining in building automation system for improving building operational performance. Energ Buildings 2014;75:109-18.
[10] Du ZM, Fan B, Jin XQ, Chi JL. Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. Build Environ 2014;73:1-11.

[11] Fan C, Xiao F, Yan CC. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Automat Constr 2015;50:81-90.
[12] Kriegel HP, Kroger P, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans Knowl Discov Data 2009;3:Article 1.
[13] Salleb-Aouissi A, Vrain C, Nortet C. QuantMiner: A genetic algorithm for mining quantitative association rules. IJCAI 2007；1035-40.
[14] Yan CC, Wang SW, Xiao F, Gao DC. A multi-level energy performance diagnosis method for energy information poor buildings. Energy 2015;83:189-203.