DEVELOPMENT OF DATA MINING-BASED BIG DATA ANALYSIS METHODOLOGIES FOR BUILDING ENERGY MANAGEMENT

FAN CHENG

Ph. D

The Hong Kong Polytechnic University

2016

The Hong Kong Polytechnic University

Department of Building Services Engineering

Development of Data Mining-Based Big Data Analysis Methodologies for Building Energy Management

Fan Cheng

A thesis submitted in partial fulfillment of the requirements for the Doctor Degree of Philosophy

February 2016

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no materials previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

_____Fan Cheng_____(Name of student)

Department of Building Services Engineering

The Hong Kong Polytechnic University

Hong Kong, P.R. China

ABSTRACT

Abstract of the thesis entitled:

Development of Data Mining-Based Big Data Analysis Methodologies for Building Energy Management

Submitted by : Fan Cheng

For the degree of : Doctor of Philosophy

at The Hong Kong Polytechnic University in Feb 2016

Today's buildings are becoming not only energy intensive, but also information intensive. Building Automation Systems (BASs) are widely installed in modern buildings for automatic monitoring and control of the operation of various building services systems. BASs collect and store a huge number of sensor measurements and control signals at short time intervals. The effective utilization of the big BAS data can help to optimize and diagnose the performance of buildings so as to improve their operational performance. However, the big BAS data are not fully utilized due to the lack of advanced data analysis techniques and tools. BASs can only perform simple data analysis, such as historical data tracking, moving averages and benchmarking. Data mining (DM) is a promising solution for the knowledge discovery from massive data

sets. However, it is extremely challenging for building automation professionals to keep up with the constantly emerging sophisticated DM techniques. Meanwhile, there is a knowledge gap between building professionals and advanced data analytics. DM itself cannot tell the value or the significance of the knowledge discovered, and domain knowledge in the building field is therefore still needed to interpret and apply the knowledge discovered. This research aims to develop generic DM-based methodologies for discovering knowledge in big BAS data and applying the knowledge to building energy management, such as identifying typical and atypical operation patterns, energy performance analysis, diagnosis and optimization.

Based on a comprehensive exploration of the state-of-the-art DM techniques using case studies on the BAS data of a high-rising building in Hong Kong, the strengths and restrictions of a variety of advanced DM techniques taking into account of the characteristics of BAS data and the building operations are understood. This dissertation first presents a generic DM-based framework for knowledge discovery in massive BAS data and applications of the knowledge for building energy management. The framework consists of five phases, i.e., data pre-processing, data partitioning, knowledge discovery, post-mining and applications. The framework and the DM techniques involved at each phase are deliberately designed considering the characteristics of BAS data and the type of knowledge to be discovered. Based on the framework developed, the methodologies for discovering and applying three different types of knowledge,

including cross-sectional knowledge, temporal knowledge and graph-based knowledge, are developed, tested and evaluated using BAS data retrieved from real buildings.

BAS data are usually stored in a single two-dimensional data table, where each column represents a variable and each row is an observation consisting of the values of different variables. Cross-sectional knowledge refers to the relationships and associations between variables (i.e., different columns) without taking into account the temporal dependency. A number of DM techniques, including clustering analysis, association rule mining and decision trees, are adopted to discover cross-sectional knowledge and to improve the reliability of the knowledge discovered. Post-mining methods, which bridge the knowledge discovered by DM techniques and domain expertise, are developed to enhance the efficiency and effectiveness in knowledge selection and application. Valuable knowledge has been discovered to understand building operation behaviors and spot energy conservation opportunities.

Different from cross-sectional knowledge, temporal knowledge discovery focuses on discovering the temporal relationships between observations (i.e., different rows). In this case, the observations are considered as multivariate time series. The symbolic aggregation approximation (SAX), motif discovery and temporal association rule mining are applied as the main DM techniques. Two post-mining methods are developed to effectively utilize the knowledge discovered. The knowledge discovered can be used to characterize the dynamics

in building operations and facilitate fault diagnosis and control optimization.

A graph-based DM technique is developed for mining BAS data with potentially complex structures, rather than just a single two-dimensional data table. It ensures the knowledge discovery efficiency when the BAS data structure is complex, e.g., data are stored in multi-relational databases and cannot be easily merged into a single data table. With the population of building information modelling, a huge amount of valuable information related to building design and operations is becoming available for analysis, such as the text data for building construction and maintenance, and the spatial information of system components. Graphs provide great flexibility in integrating and representing various types of information; and the knowledge discovered using graph-based DM is highly interpretable. The frequent subgraph mining (FSM) and graph-based anomaly detection (GBAD) are selected as the primary mining techniques. Two problems are specifically addressed, i.e., graph generation based on BAS data and efficiency enhancement in knowledge selection and application. The graph-based mining methodology has been applied to represent different types of information, based on which frequent and atypical building operation patterns are detected.

BAS data retrieved from the tallest building in Hong Kong and the Zero-Carbon Building are used to test and evaluate the methodologies. The results show that the knowledge discovered is valuable to identify dynamics, patterns and anomalies in building operations, assess building system performance and spot opportunities in energy conservation. The framework and

the methodologies can contribute to develop more powerful and sophisticated BAS tools for effective utilizing the big BAS data for building energy management.

PUBLICATIONS ARISING FROM THIS THESIS

Journal Papers Published

2015 Fan C., Xiao F., Yan C.C., A framework for knowledge discovery in massive building automation data and its application in building diagnostics, Automation in Construction 50 (2015) 81-90.

2015 Fan C., Xiao F., Madsen H., Wang D., Temporal knowledge discovery in big BAS data for building energy management, Energy and Buildings 109 (2015) 75-89.

2014 Fan C., Xiao F., Wang S.W., Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Applied Energy 127 (2014) 1-10.

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

2014 Xiao F., Fan C., Wang, S.W., Building system performance diagnosis and optimization based on data mining techniques, Chinese Journal 65 (2) (2014) 181-187.

Journal Papers Submitted

2016 Fan C., Wang D., Yan C.C., Xiao F., A pilot study on knowledge discovery in building operational data using graph-based data mining techniques,

Submitted to Energy and Buildings.

2016 Fan C., Xiao F., Yan C.C., Wang S.W., Research and applications of data mining techniques at the building operation stage – a review, Submitted to Energy and Buildings.

Conference papers

2015 Yuan Y., Fan C., Wang D., Xiao F., Developing associations between building occupancy and traffic congestion, The ACM/IEEE 6th International Conference on Cyber-Physical Systems, New York, USA, April 14-16, 2015.

2014 Fan C., Xiao F., Wang S.W., Rare event analysis of high dimensional building operational data using data mining techniques, The 3rd International High Performance Buildings Conference, Purdue University, USA, July 14-17, 2014.

2013 Fan C., Xiao F., Wang S.W., Prediction of chiller power consumption using time series analysis and artificial neural networks, CLIMA 2013, Prague, Czech, June 16-19, 2013.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisors Dr. Linda Fu Xiao and co-supervisor Prof. Shengwei Wang for their continuous support, patience, aspiring guidance and invaluable suggestions during my Ph.D study. I am very fortunate to have them as my supervisors. What I have learnt from them will not only benefit my future academic research but also my personal life.

My sincere gratitude also goes to my fellow colleagues and friends. Their kind help and support motivated me to accomplish the study. I will always cherish the companionship developed through these memorable years.

Lastly, I would like to thank my family: my parents and my wife for their trust and understanding. This journey would never commence without the encouragement and advices from my parents. The unconditional support and love I received from my wife has motivated me to be a better man both academically and personally. This journey has been amazing and I am thankful to have their company.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Motivation

Buildings have become one of the largest energy consumers around the world. According to the statistics provided by the Internal Energy Agency (IEA), buildings account for 32% of the total final energy consumption and around 40% of the primary energy consumption in most IEA countries [IEA 2015]. In Hong Kong, buildings contribute to over 90% of the total electric energy consumption and approximately 60% of the greenhouse gas emission [EMSD 2014]. To achieve the goal of sustainable development and environmental conservation, the enhancement in building energy efficiency is urgently needed.

Buildings consume energy in their whole lifecycles. The building operation stage is the most energy intensive one, as it typically accounts for 80-90% of the total energy use during building lifecycles [Ramesh 2010]. In practice, the mismatch between the intended performance at the design stage and the actual performance at the operation stage of various building services systems is a widely existing problem. Such mismatch may be caused by various reasons, such as unreliable control strategies, faults in building operations, and degradation of system components. Advanced technologies have been developed to achieve high building operational

performance. The Building Automation System (BAS) is a prominent example which integrates technologies from information science, computer science, control theory and etc. It enables buildings to be more intelligent and energy efficient by providing real-time monitoring and controls over the operations of various building services systems, e.g., the Heating, Ventilation, and Air-conditioning (HVAC) system, vertical transportation system, lighting system, and security system. In essence, BAS is a network consisting of a range of hardware devices (e.g., servers, workstations, sensors and digital controllers) and software (e.g., energy management programs and network communication protocols). A typical BAS has the ability to collect thousands of sensor measurements or control signals at short time intervals (e.g., from tens of seconds to several minutes). As a result, buildings are becoming not only energy intensive, but also information intensive.

Building operational data in BAS is typical big data. For example, in the International Commerce Centre (ICC), the tallest building in Hong Kong, at one typical office floor, over 750 sensor measurements (temperature, flow rate, pressure, humidity, power, etc.) and control signals (pump and fan speeds, valve and damper positions, sequencing signals, etc.) are collected and stored at intervals of 1 minute in its BAS. Considering only the 90 office floors in ICC, the BAS has the ability to store over 1 million pieces of data in one day, over 32 million pieces of data in one month and nearly 400 million pieces of data in one year. The volume of the stored data keeps rising over time with the building operation. Except for some critical power

consumption data, the majority of these massive BAS data sets are just left in the database, along with its contained valuable information about the actual building operations. The effective utilization of the vast amounts of building operational data available in BASs can bring significant benefits in understanding the building actual operation behaviors, evaluating operational performance, and spot opportunities in energy saving. However, the current utilization of such big BAS data is rather limited. The reason behind is twofold. Firstly, conventional data analytics adopted in the building automation industry, which usually rely on domain expertise, physical principles and statistics, are neither efficient nor effective in handling massive amounts of data. Meanwhile, it is extremely challenging for building automation professionals to keep up with the constantly emerging sophisticated big data analysis techniques. The knowledge gap between building professionals and data scientists significantly hinders the utilization of the big BAS data. Secondly, the BAS data are usually complex and have poor quality. The intrinsic complexity in BAS data stems from the dynamic operations of various building equipment under changing conditions. The BAS data usually contains a substantial number of outliers and missing values due to the widespread errors in data sampling, transmission and storage processes. The knowledge discovered from low-quality data can be hardly transformed into meaningful and actionable measures. The building industry's desire for effective and convenient big data analysis techniques for analyzing the massive BAS data has become stronger and stronger.

Today's explosive growth of information greatly promotes the development of big data analysis technology in various industries. Big data analytics has been attracting increasing attention from both academic communities and industries. A typical approach to utilizing big data is to mine the data in order to discover the hidden knowledge, in the forms of patterns, correlations, associations, classification, regressions, and etc. The development of big data analytics in the building industry is still at its beginning stage. There are a lot of unknowns and uncertainties on the development and applications of advanced big data analytics for analyzing big BAS data. People are excited about big data and big data analytics, but few of them clearly know how and to what extent the industry can benefit from big data. Therefore, this research primarily aims to bridge the building industry and big data analytics using a multidisciplinary approach and bring innovative ideas and enabling technology for effective utilizing the massive amounts of building operational data in BAS. Based on comprehensive review of advanced data analytics and in-depth understanding of building operations, this research develops generic solutions for analyzing big BAS data and maximizes their practical values in building energy management for improving building operational performance.

1.2 Aim and Objectives

This research aims to develop a generic DM-based framework and associated

specific DM-based methodologies for knowledge discovery in the big BAS data as well as their applications in building energy management. The aim can be accomplished by addressing the following major objectives:

1. To explore the state-of-the-art DM techniques using case studies on the BAS data of a high-rising building. The strengths and restrictions of a variety of DM techniques in analyzing big BAS data will be identified. Typical types of knowledge which can be discovered through mining big BAS data will be analyzed for applications in building energy management.
2. To develop a generic framework for mining big BAS data, considering the unique BAS data characteristics and the actual needs of building professionals.
3. To develop DM-based methodologies, based on the generic framework developed, for discovering cross-sectional knowledge, temporal knowledge and graph-based knowledge from big BAS data.
4. To apply the DM-based methodologies to BAS data from real buildings and give recommendations on using the methodologies in practical building energy management.

1.3 Thesis Organization

The whole thesis is divided into 7 chapters. The main content of each chapter is presented as follows.

Chapter 1 presents the motivation of this research, the aim and objectives, and the organization of this thesis.

Chapter 2 presents a literature review on the conventional methods for building energy management, and DM-related research and applications at three stages of the building lifecycle, including the building design, construction and operation stages.

Chapter 3 introduces the generic DM-based framework developed and the research facilities, including the buildings and the BAS data to be mined and the computation tools used.

Chapter 4 develops a methodology to discover the cross-sectional knowledge in BAS data. The methods for data preprocessing, knowledge discovery and post-mining are presented in detail. The methodology is validated using the real-world BAS data and the knowledge discovered is presented.

Chapter 5 develops a methodology to discover the temporal knowledge in BAS data. It is specifically designed to discover the frequent sequential patterns and the temporal associations in BAS data. The methodology is validated using the real-world BAS data and the knowledge discovered is presented.

Chapter 6 develops a methodology for the knowledge discovery in BAS data with potentially complex data structures. It specifically addresses the challenges of

knowledge discovery from BAS data stored in multi-relational data. The methodology is validated using the real-world data and the knowledge discovered is presented.

Chapter 7 summarizes the work presented in this thesis, and gives some recommendations for future research and applications.

CHAPTER 2 LITERATURE REVIEW

Building energy efficiency has become one of the top concerns in the building industry, especially the energy efficiency at the building operation stage, as it accounts for 80-90% of the total building energy consumption in building lifecycle [Dalene 2012]. In the past decades, both the academics and building professionals are seeking for the solutions to enhance the building energy efficiency throughout the building lifecycle. One essential approach is to perform analysis based on the information collected. With the advance in information technologies, a large amount of information is being collected throughout the building lifecycle. Conventional analytics, which primarily rely on building physics and domain expertise, usually lack of efficiency and effectiveness when handling massive datasets. By contrasts, data mining (DM) is a highly promising technology which can efficiently and effectively extract the hidden knowledge from large data. It could be a powerful tool to tackle the challenges brought by the big data era.

DM techniques have been adopted in previous research to perform different tasks. This chapter presents a comprehensive review on the research and applications of DM techniques in the building industry, especially at the building operation stage. Section 2.1 briefly introduces the conventional methods adopted at the building operation stage for building energy management. Section 2.2 introduces the background of DM technology. Section 2.3 reviews the DM research and applications for building

operation management. Section 2.4 presents the DM applications at the other two stages in building life cycle, i.e., building design and construction stages. The summary is given in section 2.5.

2.1 Conventional Methods for Building Energy Management

Building energy management has been a hot research topic for the last few decades. Conventional methods mainly adopt engineering expertise, physical principles and statistics to analyze the building operational data. They are mainly used for handling two typical tasks, i.e., fault detection and diagnosis (FDD) and control optimizations. The following contents serve as a brief review on these topics. A more detailed review can be found in [Wand and Ma 2008; Ma and Wang 2009; Xiao and Wang 2009; Katipamula and Brambley 2005 (a); Katipamula and Brambley 2005 (b)].

The FDD is typically applied at three levels, i.e., building level, system level, and component level. Through literature review, it is found that the FDD at the building level mainly targets at the building energy consumption [Wu and Sun 2011]. The FDD at the system level considers the interactions between subsystems and components [Zhao et al. 2013]. The FDD at the component level is mainly relied on the physical relationships among parameters [Xiao et al. 2014]. In addition, the FDD research on sensors [Wang and Xiao 2004; Xiao et al. 2009] is another hot topic and

has been closely associated with the research mentioned above. For instance, Wang et al. proposed a FDD methodology for the HVAC system, which integrated concepts of system level FDD and sensor FDD [Wang et al. 2010]. This study took into account the influence of sensor faults on system level FDD. Firstly, the principal component analysis was applied to identify the bias in sensor measurements, based on which regression methods were applied to detect system level faults. The methods used in FDD can be categorized into two types, i.e., model-based and data-driven methods. The target variables of FDD can be summarized into three categories, i.e., the energy consumption (e.g., cooling load and electricity consumption [Zhao and Magoules 2012]), performance indices (e.g., chiller COP [Lee and Lu 2010]), and physical parameters (e.g., indoor temperature [Kruger and Givoni 2008]). The model-based methods can be further classified into quantitative and qualitative methods [Katipamula and Brambley 2005(a); Katipamula and Brambley 2005(b)]. Quantitative model-based methods mainly adopt domain expertise and physical principles to build physical models [Bendapudi and Braun 2002]. Qualitative methods, by contrast, are in the forms of expert systems [House et al. 2001]. In general, the model-based methods can yield fairly good performance. However, the models developed can be of high complexity and the actual usefulness is largely restricted by the data availability. The data-driven methods mainly adopt grey-box [Jia and Reddy 2003] and black-box models [Reddy et al. 2003] to perform FDD. Domain expertise and historical data are the two pillars for model development. The inputs to the model are mainly selected

based on domain expertise. The historical data are then applied to fit the model. Compared to the physical principles-based methods, the data-driven methods have fewer constraints in real implementations. They are less restricted by the assumptions of physical principles and can capture the unexplainable variations in system operations.

The control optimization task is mainly focused on the building HVAC system. Previous research can be summarized at two levels, i.e., system and component levels. The control optimization at the system level aims to perform global optimization. For instance, [Lu et al. 2005] proposed an optimization strategy which considers the interactions between the water- and air-sides of the HVAC system. It firstly used mathematical methods to model the building cooling load and electricity consumption. Evolutionary algorithms were then used to optimize the controllable variables, such as the supplied chilled water temperature and the chilled water flowrate. The objective function was formulated in such a manner that both the energy consumption and the indoor comforts were taken into account. The control optimization at the component level mainly focuses the operational performance of individual components and therefore, can be regarded as a local optimization problem. Typical optimization targets include chillers (e.g., chiller sequencing control and the optimization of supplied chilled water temperature) [Sun et al. 2009; Wang et al. 2010], variable-speed pumps [Ma and Wang 2009; Wang and Ma 2010] and etc. The optimization methods can be classified into linear and non-linear types. We direct

interested readers to [Wang and Ma 2009] for a detailed summary.

As abovementioned, the conventional approaches adopted in building operational performance management usually obtain knowledge as quantitative and qualitative models, expert rules, and statistics. Such knowledge representations are very similar to those obtained by DM techniques. DM technology is more flexible and efficient when the data is of massive volume and high complexity. It can better utilize the vast amounts of building data and effectively discover novel knowledge in various representations. This is also the main motivation of this research.

2.2 Data Mining Technology

2.2.1 Overview

Nowadays, overwhelming amounts of data are being generated, collected and stored worldwide. According to the International Data Corporation, 2.8 trillion gigabytes of data were generated in 2012, and this figure is expected to reach 40 trillion gigabytes by 2020 [Gants and Reinsel 2012]. Despite of the huge data amount, only a limited proportion is utilized for detailed analyses. For instance, out of the 2.8 trillion gigabytes data produced in the year of 2012, only 0.5% of them are properly used for analysis [Gants and Reinsel 2012]. A common challenge faced by many industries is being “information rich, but knowledge poor”, and the main reason is due to the lack of advanced analytics for processing large data sets [Han and Kamber 2011].

Data mining (DM) arises as a solution to the big data challenge. DM, or knowledge discovery from databases (KDD), is defined as a nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [Fayyad et al. 1996]. Currently, DM has been successfully used in various industries. For instance, in the retailing industry, DM is used to mine the customer purchasing behavior. The extracted knowledge can be applied to benefit cross-selling [Olson and Delen 2008]. In the banking and financial industry, DM is used to identify customer values, develop revenue maximization programs, and detect credit card frauds [Olson and Delen 2008]. In the healthcare industry, DM helps insurers to detect fraud, physicians to find effective treatments, and pharmaceutist to develop new products [Koh and Tan 2005]. Other applications include customer relationship management [Ngai et al. 2008], human resource management [Strohmeier and Piazza 2013], counterterrorism [DeRosa 2004], and epidemic detection [Ginsberg et al. 2009].

In general, the DM process consists of seven steps, i.e., data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and application [Han and Kamber 2011]. The first four steps aims to prepare a suitable data set for mining and can be merged as the data preprocessing step. Data preprocessing can be very time-consuming. It is estimated that data preprocessing normally accounts for more than 80% of the total analysis time [Zhang et al. 2003]. The data mining step refers to the implementation of various analytical techniques

and algorithms to discover knowledge. The pattern evaluation step is about post-mining with the aim of evaluating the mining process, interpreting the mining results, selecting interesting or potentially useful knowledge. The last step is application, which transform the knowledge obtained into actionable measures.

2.2.2 Data Mining Techniques

DM techniques can be roughly classified into two categories, i.e., supervised and unsupervised methods. Supervised methods need to classify the data into inputs (i.e. independent variables) and outputs (i.e., dependent variables) and aim to establish the relationships between the inputs and the outputs by the use of knowledge discovered. Supervised methods can be used for either regression or classification. Popular methods include the Naïve Bayes [Rish 2001], adaptive Bayes network [Acar et al. 2007], support vector machines [Cortes and Vapnik 1995], artificial neural network [Gershenson 2003], decision trees [Apte and Weiss 1997], random forests [Breiman 2001], and ensemble learning [Dietterich 2001].

By contrast, unsupervised methods focus on discovering the intrinsic data structure, correlations, or associations from big data. Popular techniques include the clustering analysis and association rule mining (ARM). Clustering analysis aims to partition the data into several clusters or groups based on data similarity. The similarity between data sets can be evaluated by either distance or density. Various

metrics, such as the Minkowski distance for numeric values and the Hamming distance for categorical values, are used for similarity evaluation. Clustering algorithms can be classified into hierarchical-based, partitioning-based, density-based, grid-based, constraint-based, and subspace methods [Grira et al. 2005]. ARM derives rules specifying the associations among variables. It can be further divided into quantitative ARM and qualitative ARM based on whether the data are numeric or categorical. Popular algorithms include the Apriori [Agrawal and Srikant 1994], FP-growth [Han et al. 1999], GAR [Meta and Alvarez 2002], QuantMiner [Salleb-Aouissi et al. 2013], and etc.

2.2.3 Data Mining Software

A larger number of software tools are available to perform DM. The most widely used DM software includes RapidMiner, R, Weka, Python, SAS, MATLAB, Statistica, IBM SPSS, KNIME, Orange, and etc. [KDnugget 2014].

Potential users may select a suitable tool considering the costs, user-friendliness, and desired DM tasks to be performed. The costs refer to whether the software tool is commercial or open-source. Commercial software usually has better user support and the validity of the mining result is guaranteed. By contrast, open-source software is free, and newly developed algorithms are more likely to be implemented in open-source software first.

Software with graphical user interface (GUI) is much easier to use, especially for the users with little programming knowledge. However, the flexibility provided during the mining process is usually restricted. To use software with command line interface (CLI), users may need to master a new programming language and the learning process can be time-consuming.

Lastly, users may choose the software considering the DM tasks to be performed. Some software, such as the SAS Enterprise Miner, provides an integrated package for nearly all kinds of DM-based analysis. Others may only focus on certain types of DM tasks. For instance, the CART 5.0 Decision Tree only provides decision tree-based solutions for predictive modeling. A more detailed review on data mining software can be found in [Mikut and Reischl 2011].

2.3 DM Research and Applications for Building Operation Management

Building operation has drawn particular attention from both the academic and industrial worlds. It accounts for 80% to 90% of the total building green house gas emission, and is directly linked to occupant comforts and the realization of building functionality [Dalene 2012].

One prominent problem throughout building lifecycle is the mismatch between design and actual performances. Such mismatch may due to various reasons, such as

all kinds of operation faults, improper control strategies, or performance degradation. By analyzing the building operational data, it is possible to find useful information to enhance the building operational performance. However, conventional analysis methods may not cope well with the ever-increasing amount of building operational data. The DM's excellent capability in knowledge extraction makes it very promising in utilizing the massive building operational data. DM techniques have been applied to facilitate the on-going commissioning process, which typically performs tasks such as benchmarking, performance tracking, and fault detection and diagnosis [Choiniere and Corsi 2003; Djuric and Novakovic 2009; Ginestet and Marchio 2010; Ahmed, et al. 2013; Ginestet et al. 2013]. As shown in Figure 2.1, this section reviews the relevant research from three perspectives, i.e., predictive modeling, fault detection and diagnosis, and control optimization.

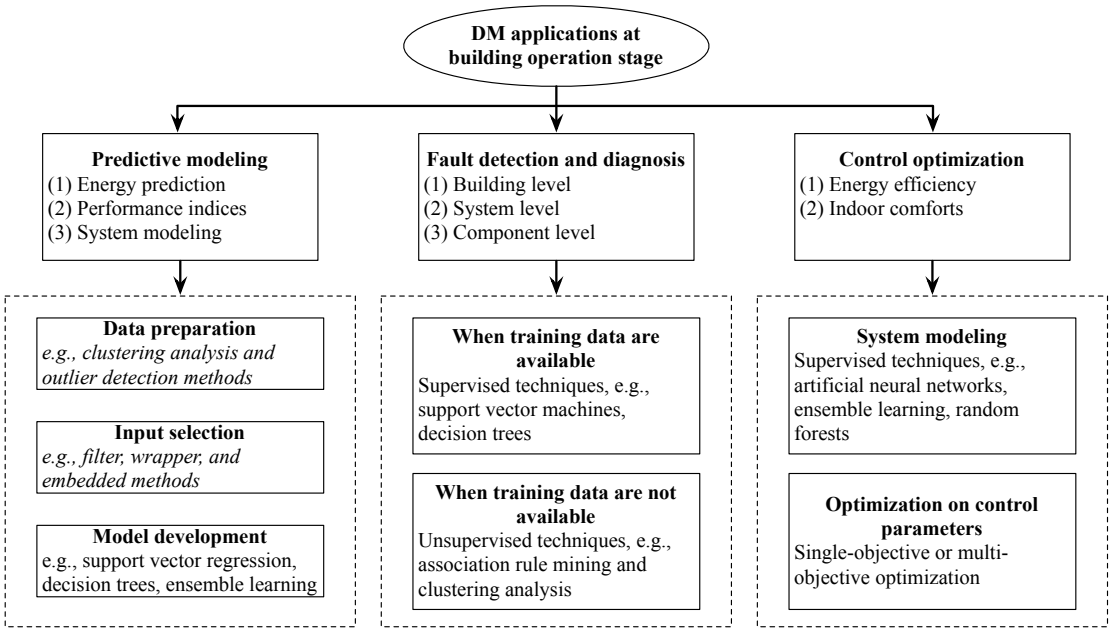


Figure 2.1 DM applications at building operation stage

2.3.1 Predictive Modeling

DM techniques have been used for the predictive modeling of various variables, including the cooling or heating load [Ben-Nakhi and Mahmoud 2004; Hou et al. 2006; Li et al. 2009; Kusiak et al. 2010; Kwok et al. 2011], energy consumption [Azadeh et al. 2008; Kusiak et al. 2010; Nagi et al. 2011; Mena et al. 2014; Ahmad et al. 2014], indoor environment [Ruano et al. 2006; Wu and Clements-Croome 2007; Mustafaraj et al. 2010; Kim et al. 2013], and performance indices [Kucuksille et al. 2009; Yu et al. 2010; Kucuksille et al. 2011; Saez et al. 2013; Chou et al. 2014].

Researchers have been working towards a more systematic process of predictive modeling, especially in the areas of data preparation, input selection, and model development. At the early stage, main effort was put on investigating the use of advanced DM algorithms for model development and the technical issues in optimizing model parameters. Afterwards, researchers began to explore the potential of DM techniques in other steps, such as data preparation and input selection.

Model Development

At the early stage, DM techniques were mainly used as a substitute for conventional methods (e.g., linear regression). Encouraging results have been obtained, as DM techniques are better in capturing complex and nonlinear relationships. Popular tools include support vector machine (SVM) [Dong et al. 2005;

Zhao and Magoule 2010], artificial neural network (ANN) [Ben-Nakhi and Mahmoud 2004; Yang et al. 2005; Karatasou et al. 2006; Kusiak et al. 2010], decision tree [Tso and Yau 2007; Yu et al. 2010], ensemble learning [Jetcheva et al. 2014], and time series analysis [Yun et al. 2012; Ogunsola et al. 2014].

It is worth mentioning that domain knowledge still plays the dominant role in other steps, e.g., data preparation and input selection. For instance, Dong, Cao, and Lee first applied support vector machine (SVM) to predict the monthly electricity consumption [Dong et al. 2005]. This study investigated the use of SVM in modeling building energy consumption. It mainly focused on the technical issues of developing SVM models, such as model parameter setting and kernel selection. The prediction results were satisfactory and validated the use of SVM in energy prediction. In this study, only three input variables, i.e., monthly mean temperature, relative humidity, and global solar radiation, were used for prediction and they were selected based on domain knowledge.

Similarly, Li et al. applied SVM to predict the hourly cooling load in commercial buildings [Li et al. 2009]. The results validated that SVM was effective in cooling load prediction. Again, domain knowledge was applied to select input variables, i.e., outdoor temperature, humidity and solar radiation.

Ben-Nakhi and Mahmoud adopted general regression neural networks (GRNN) to predict the next-day cooling load [Ben-Nakhi and Mahmoud 2004]. A parametric study was carried out to determine the optimum parameters of GRNN. Based on

domain knowledge, the external outdoor temperature records for the last 24-hour were selected as inputs. The results showed that neural networks could be very powerful in load prediction.

Yu et al. developed a decision tree method for classifying the building energy demand [Yu et al. 2010]. One particular advantage of the decision tree method is that a flowchart-like tree structure can be generated for better interpretation. The results showed the developed model could achieved accurate prediction on the building energy use index (EUI). In addition, the model could automatically identify and rank significant variables to the output. In this study, ten input variables were selected based on domain knowledge.

The above-mentioned studies validated the use of advanced DM algorithms in model development. However, the potential of DM techniques was not fully realized. One particular problem is that the input variables are predefined based on domain knowledge. In such a case, the developed method may not obtain the optimal performance for different buildings, due to the variation in operation and data availability.

Input Selection

To tackle the above-mentioned challenge, researchers began to extend the use of DM techniques in input selection. Input selection helps to improve the prediction accuracy, reduce the computation load, and gain new insight into the underlying

process [Guyon and Elisseeff 2003]. Through literature review, it is found that three types of input selection methods are primarily used in the building field, i.e., filter, wrapper, and embedded methods. Filter method evaluates the importance of input variables based on statistical measures. Univariate techniques, such as the correlation coefficient, chi-squared test, and information gain, are commonly used. It can be regarded as a pre-processing step, and the selection is independent with the predictive algorithm.

For instance, Kusiak, Li, and Zhang applied correlation method to facilitate the input selection process [Kusiak et al. 2010]. Combining with the boosting algorithm, 2 out of 13 variables were selected as inputs to predict the building steam load. The input selection method successfully removed redundant and irrelevant variables. The results indicated that the prediction performance was guaranteed while the computational load was reduced.

Similarly, Chou, Hsu, and Lin adopted the Pearson correlation to select inputs for predicting chiller COP [Chou et al. 2014]. Various predictive models, including the ANN, SVM, classification and regression tree (CART), chi-squared automatic interaction detector (CHAID), were constructed based on the selected inputs. The research results showed that accurate predictions could be achieved and the predictions could be used for fault detection.

Mena et al. used the two filter methods for input selection [Mena et al. 2014]. Correlation method was used to measure the linear dependency, while mutual

information was applied to measure the nonlinear one. The method identified 5 out of 20 variables as model inputs. The developed model could accurately predict the building electricity consumption.

Zhao and Magoules also discussed the use of filter methods in enhancing the performance of building energy prediction [Zhao and Magoules 2012]. The effectiveness of two filter methods, i.e., correlation coefficient and regression, gradient-guided feature selection, were investigated. SVR models were then developed to predict building energy consumption. The results showed that the inclusion of variable selection was effective in improving the prediction accuracy while reducing the computational time.

To summarize, filter methods are computational efficient and scalable to high-dimensional data. However, the variables selected tend to be redundant and the interaction with predictive algorithms is ignored. Therefore, the resulting prediction performance may not be optimal compared to other feature selection methods.

The second is the wrapper method, which integrates the model selection within the input selection process. For each input subset, the prediction performance is evaluated based on a certain predictive algorithms. Popular wrapper methods include forward selection, backward elimination, and genetic algorithm [Saeys et al. 2007]. Kusiak, Tang, and Xu adopted the wrapper method to select inputs for prediction models in HVAC system [Kusiak et al. 2011]. Three algorithms, i.e., greedy, linear forward, and genetic algorithm, were used to search the optimal input subset. The

performance of each input subset was evaluated considering four predictive algorithms, i.e., linear regression, pace regression, SVM, and multi-layer perceptron (MLP). The method was reported to be effective in selecting compact input subset.

Kolter and Ferreira presented a data-driven approach to model the building energy consumption in residential and commercial buildings [Kolter and Ferreira 2011]. One wrapper method, i.e., greedy forward selection, was used to for input selection. It was recommended to integrate the input selection process for the sake of model interpretability and computational load reduction.

The main advantage of the wrapper method is that it considers the dependency between input variables and the interaction between input variables and predictive algorithms. However, it may suffer from the risk of overfitting and can be very computationally expensive [Saeys et al. 2007].

The third method, i.e., embedded method, performs the variable selection during the process of model training and is specific to a given predictive algorithm. Popular methods include weighted naïve Bayes and the variable selection using the weight vector of SVM. Similar to the wrapper method, it considers the dependency between inputs and the interaction with predictive algorithms. One particular advantage is that embedded method is far less computationally expensive [Saeys et al. 2007]. Fan, Xiao, and Wang applied the embedded method to better predict the next-day building electricity consumption and peak power demand [Fan et al. 2014]. The recursive feature elimination (RFE) algorithm was used to select the optimal inputs considering

different predictive algorithms, e.g., ANN, SVR, and random forests. The research results showed that RFE could improve the prediction accuracy while reducing the computational load.

Data Preparation

DM techniques have also been used for data preparation, particularly for data partitioning. Building operational data can be very complex, containing observations under different operation conditions. Therefore, it may not be wise to using the whole data to train one universal model. Data partitioning helps to group observations according to their similarities. Consequently, the observations in the same cluster tend to have similar operation conditions. Based on these clusters, individual model can then be developed to achieve more accurate predictions. One specific DM technique, i.e., clustering analysis, is well fitted to this type of tasks. Clustering analysis groups the data into several clusters, with the aim of maximizing the intra-cluster similarity while minimizing the inter-cluster similarity. It can be used to as a pre-processing step, with the aim of identifying typical operation patterns and detecting abnormal observations. Tang, Kusiak, and Wei utilized clustering analysis to facilitate the predictive model development [Tang et al. 2014]. The authors applied *k*-means algorithm to obtain different clusters of data, based on which individual predictive models were then developed. It was claimed that using clustering analysis to preprocess the data was able to decrease the prediction errors and the computational

load.

Jota et al. develop a 4-step methodology to forecast the building electricity consumption [Jota et al. 2011]. Hierarchical clustering was applied to identify the number of typical load curves. Then, mathematical models were developed to represent typical load curves. The last two steps were designed to predict the accumulate energy consumption and the maximum daily demand, respectively. The research results showed that the proposed method could simply and quickly predict the energy consumption and maximal demand. The obtained knowledge could be used for energy managers to identify anomalies, manage energy costs, and automate demand response strategies.

Jain and Satish proposed a novel clustering-based short-term load forecasting method [Jain and Satish 2009]. Prediction was made for the next 24 hours with a time resolution of 30 minutes. Clustering analysis was applied as a preprocessing step to improve the prediction performance. SVM models were developed for each day of the week, taking into account the results of clustering analysis. By comparison, it showed that significant improvement could be achieved when clustering was integrated into the predictive modeling process.

2.3.2 Fault Detection and Diagnosis

Fault detection and diagnosis (FDD) has been a hot topic in the building field for

decades. Conventional methods mainly adopt physical principles [Norford et al. 2002; Castro 2002] and domain expertise [House et al. 2001; Schein et al. 2006] to detect and diagnose faults in building operations. The rapid growth in the amount of building operational data has made it less efficient to detect and diagnose faults through conventional strategies. The availability of massive building operational data has provided another approach to FDD, i.e., the data-driven approach. Some research has been done to explore the usefulness of DM techniques in such area. Basically, the DM-related research can be organized according to the criteria of whether training data are available or not. With the presence of training data, supervised techniques can be used for detection and diagnosis. Otherwise, unsupervised techniques are used to examine the intrinsic data structure, correlations, or associations, based on which faults are detected or diagnosed.

FDD Based on Supervised DM Techniques

Supervised DM techniques have been widely used in predictive modeling owing to their excellent ability in mapping complex and nonlinear relationships. When training data are available, one commonly used approach for fault detection is to model the normal conditions first and then, detect potential faults through comparison. Supervised DM techniques have been widely used for the modeling purpose. Fault diagnosis can be performed using supervised learning methods when the training data contain both normal and faulty observations. In such a case, fault diagnosis is

transformed into a classification problem. Popular tools include the artificial neural networks (ANN) [Lee et al. 2004; Magoules et al. 2013], support vector machines (SVM) [Liang and Du 2007; Zhao et al. 2013], principal component analysis (PCA) [Wang and Cui 2005; Wang and Xiao 2006; Jin and Du 2006;], fisher discriminant analysis (FDA) [Du et al. 2007; Du and Jin 2008], and wavelet analysis [Chen et al. 2006; Du et al. 2009].

PCA methods are mainly used for sensor FDD. Wang and Xiao presented a principal component analysis (PCA)-based strategy for AHU sensor fault detection and diagnosis [Wang and Xiao 2004]. Sensor faults were detected using the Q-statistic and diagnosed using the Q-contribution plot. Several simple physical rules were integrated to enhance the fault isolation ability of the PCA method. The robustness of the proposed method was verified through simulation and site data. Compared to the training of neural networks or other black-box models, PCA-based strategy was much easier to implement. In addition, the fault isolation ability of PCA models could be further enhanced with physical reasoning.

Hou et al. applied rough set and ANN to detect and diagnose sensor faults in building air conditioning system [Hou et al. 2006]. The proposed method was tested using an existing HVAC system in China. The sensor faults of supplied and returned chilled water temperature were successfully detected and diagnosed. It was claimed DM techniques could provide a good means of generating useful residuals for sensor fault detection and diagnosis.

Hu et al. presented a self-adaptive chiller sensor fault detection strategy based on PCA [Hu et al. 2012]. The proposed method was specifically design to enhance the detection efficiency at low sensor fault level. The usefulness of the proposed method was validated by the operational data of a screw chiller system. Compared to conventional PCA-based strategy, the proposed method was able to remove the error samples with a self-adaptive loop in the process of PCA model development and thereby, enhancing the fault detection efficiency.

PCA-based methods have also been used for the FDD of system or component operations. Wen and Li applied the PCA and pattern matching methods for the FDD of air handling units [Wen and Li 2014]. For each new observation, the pattern matching method identified the similar operation conditions in the historical data. PCA model was then built on this data subset. The Q-residual was used to detect the fault, if any. The method was validated using the ASHRAE 1312-RP and 1020-RP data, indicating that the sensitivity of fault detection in AHU system could be enhanced significantly.

Artificial intelligence (e.g., ANN) and machine learning methods (e.g., SVM, SVR) are mainly used for the FDD of system operations. Bailey and Kreider developed an automated chiller fault detection diagnostics tool using neural networks [Bailey and Kreider 2003]. The tool was able to classify the current state of chillers given a vector of observations. The model development required the availability of empirical data containing both normal and abnormal observations. The proposed

method was validated by extensive experiments and the performance was satisfactory. Similarly, Tassou and Grace utilized artificial intelligence to detect and diagnose faults in vapour compression refrigeration systems [Tassou and Grace 2005]. Fault detection was based on the comparison between predicted fault-free values and actual values of ten parameters. ANN models were developed for the prediction. An expert system was used to diagnose the faults. The method was capable of distinguish faulty and fault-free conditions, steady state and transient operations, refrigerant leakage and overcharge conditions. Capozzoli, Lauro, and Khan investigated the usefulness of artificial ensembling networks in automatic fault detection [Capozzoli et al. 2015]. The ANN ensembles were trained to capture the relationship under normal conditions. The model residuals were analyzed using the peak detection and the GESD methods. Compared to statistical methods, the proposed method could successfully identify outliers under boundary conditions and provide more robust results.

Another example of using ANN can be found in [Zhou et al. 2009]. The authors adopted fuzzy modeling and ANN techniques for the FDD of centrifugal chillers. The performance indexes under normal conditions were modeled using regression analysis. The residuals between the model estimates and normal data were used for fault diagnosis. Fuzzy model was used to deduce a quantitative diagnostic classifier. ANN was applied for fault identification. The strategy was validated using the ASHRAE 1043-RP data. Magoules, Zhao, and Elizondo (2013) demonstrated the use of recursive deterministic perceptron (RDP) neural network in detecting anomalies in

building electricity consumption [Magoules et al. 2013]. *EnergyPlus* was used to simulate both normal and abnormal energy consumption data for various components, including fans, pumps, and chillers. RDP neural networks were then developed to detect faulty conditions. Afterwards, a fault diagnosis scheme was designed and it was able to output the potential sources for a given faulty sample in a descending order of possibilities. The effectiveness of the method was validated through experiments.

SVM or SVR methods are also popular tools. Yan et al. developed a robust FDD strategy for chiller FDD using time series modeling and machine learning techniques [Yan et al. 2014]. Time series modeling was used to preprocess the data and SVM was applied to detect the changes in model parameters. The proposed method was validated using the ASHRAE 1043-RP data. It was shown that five typical chiller faults could be accurately identified with low false alarm rates. Zhao, Wang, and Xiao devised a novel FDD method for centrifugal chillers [Zhao et al. 2013]. A new performance index, i.e., the heat transfer efficiency of sub-cooling section, was proposed. SVR models were developed to calculate the performance indexes under normal conditions. The exponentially-weighted moving average (EWMA) control charts were used for fault detection. A rule table was proposed for fault diagnosis. The method was validated by the experimental data from the ASHRAE 1043-RP, considering 6 typical chiller faults. It was shown that significant improvement in FDD performance could be achieved compared to conventional method.

Some research has been done to utilize Bayesian analysis for FDD. As an example, Zhao, Xiao, and Wang developed a generic intelligent FDD strategy for chillers using Bayesian belief network [Zhao et al. 2013]. A three-layer Diagnostic Bayesian Network (DBN) was developed to capture the causal relationships between faults and symptoms. Probability analysis and graph theory were applied to calculate the posterior probabilities of faults given the observed evidence. The ASHRAE 1043-RP data were used for validation. It was claimed that the method was especially useful when handling incomplete and conflicting information. Similarly, the method was applied to the FDD of variable air volume (VAV) terminals [Xiao et al. 2014]. The causal relationships between faults and symptoms were mapped using the DBN. Two rules were used to isolate the fault. Simulation tests showed that 10 typical VAV terminal faults could be effectively detected and diagnosed.

FDD Based on Unsupervised DM Techniques

In practice, it is usually not possible to obtain a training data set with labeled observations. In such a case, unsupervised DM techniques are normally used to identify the faults or anomalies. Compared to the previous section, this area is less explored. Popular techniques include outlier detection [Seem 2007], association rule mining (ARM) [Yu et al. 2012; Cabrera and Zareipour 2013], and clustering analysis [Khan et al. 2013; Du et al. 2014; Janetzko et al. 2014; Panapakidis et al. 2014].

Through literature review, it is found that unsupervised methods are mainly used

to detect faults and anomalies in energy consumption data. Cabrera and Zareipour used association rule mining to identify the lighting energy waste patterns in educational institutes [Cabrera and Zareipour 2013]. The data set contained 7 variables, i.e., season, time, day type (i.e., weekdays or weekends), occupancy status, event, day of week (i.e., Monday to Sunday), and waste status. Association rules were derived to present the relationships between waste status and other variables. The knowledge discovered was used to regulate the lighting energy use. Simulation results showed that as high as 70% of the lighting energy could be saved.

Miller, Nagy and Schlueter developed an automated daily pattern filter to find anomalies in building operational data [Miller et al. 2015]. This research focused on mining the temporal relationship embedded in building operational data. The symbolic aggregate approximation method was applied to preprocess the time series data. The most infrequently happened data sequences, or discords, were filtered out for detailed inspection. Clustering analysis was adopted to discover the most frequent patterns, or motifs. The method was applied to two case studies and the results confirmed the capability in finding discords and motifs in time series data.

Yu et al. applied association rule mining (ARM) to discover useful knowledge about energy conservation [Yu et al 2012]. ARM was used to derive the associations and correlations between building operational data. It was recommended to use at least 2-year building operational data for comparison and inference. The derived rules were successfully used to detect equipment faults and identify energy waste

conditions.

Khan et al. adopted three DM techniques, i.e., CART, *k*-means, and DBSCAN, to detect abnormal lighting energy consumption [Khan et al. 2013]. The hourly-recorded data were firstly classified or clustered. Then, outlier detection algorithms, i.e., generalized extreme studentized deviate (GESD) and boxplot statistical method, were applied for the discovery of abnormal observations in each class or cluster. The results showed that the combine use of CART and GESD could lead to accurate fault detection. The DBSCAN algorithm could be used alone to detect potential faults, as it has the ability to group all outliers in a single cluster. However, the detection performance may be compromised. The proposed method could be used for preventive maintenance. In addition, the productivity could be enhanced as the fault detection process was automated.

Seem described a novel method for detecting abnormal energy consumption in buildings [Seem 2007]. The proposed method used the generalized extreme studentized deviates method to efficiently determine the abnormality degree of building electricity consumption. The computation was efficient and no training data was required. Field tests validated the usefulness of the method.

Jakkula and Cook compared two methods for the detection of anomalies in building electricity consumption [Jakkula and Cook 2010]. The statistical method adopted the concepts of *t*-distribution to identify outliers. The clustering method was based on the *k*-nearest neighbor algorithm and dynamic time warping. The

experiments showed that the clustering-based method was more reliable than statistical-based method.

2.3.3 Control and Optimization

Conventional and optimization methods adopt analytics [Tashtoush et al. 2005] and simulation-based methods [Lu et al. 2005a; Lu et al. 2005b] to optimize building operation. Building systems are complex, nonlinear, and containing massive amounts of variables. Therefore, conventional methods may not solve the optimization problem efficiently. This section reviews the DM-related research on control and optimization. The general approach can be summarized as follows. Predictive modeling is firstly used to construct the relationship among parameters to be optimized, constraint variables, and target variables (e.g., energy consumption). Optimization algorithms are then used to optimize the parameters according to user-specified cost function. Due to the complexity in building operation, nonlinear local techniques (e.g., direct search, sequential quadratic programming, and conjugate gradient) may not be able to find the global optimum [Wang and Ma 2008]. By contrast, evolutionary computing techniques have shown great potential in handling complex optimization problems in building operation management. Popular methods include genetic algorithm (GA) [Chang 2005; Chang et al., 2009; Beghi et al. 2011], particle-swarm optimization (PSO) [Ardakani et al. 2008; Lee and Lin 2009; Beghi et

al. 2012], differential evolution (DE) [Lee et al. 2011; Ozcan et al. 2013], and neuro-evolutionary method [Chow et al. 2002; Chen et al. 2014]. These studies can be further classified into two categories, i.e., single-objective and multi-objective optimizations.

Single-objective Optimization

Single-objective optimization has been mainly applied to optimize the performance of individual component. One typical problem is to determine chiller loadings of multi-chiller systems. The general approach is to first establish the relationship between the part-load ratio (PLR) and chiller energy consumption. Then, optimization algorithm was used to minimize the total energy consumption of chillers while meeting the cooling demand.

Chang, Lin, and Chuang employed GA to optimize the chiller loading problem [Chang et al. 2005]. The PLRs of individual chillers were encoded using binary strings. Compared to the Lagrangian method, the binary GA-based method was able to solve the convergence problem at low cooling demands. However, the optimized energy consumption might rise a little.

Since the PLRs are continuous variables, it may not be optimal to encode the PLR using binary strings. Ardakani, Ardakani and Hosseinian used continuous GA-based and PSO-based methods to solve the same problem [Ardakani et al. 2008]. The research results showed the continuous GA-based and PSO-based method could

result in better performance than the binary GA-based method. More specifically, continuous GA-based method could find more precise solution than binary GA-based method. PSO-based method converged faster than both types of GA methods. Lee and Lin compared the optimization performance of PSO-based, Lagrangian, and GA-based method [Lee and Lin 2009]. Two case studies were carried out and validated the superiority of PSO-based method. It was shown that PSO-based method could overcome the convergence problems at low cooling demands with better energy consumption solutions.

Other approaches, such as DE and firefly algorithm, have also been used for this problem. Lee, Chen, and Kao applied DE to optimize the chiller loadings. The proposed method was tested in two multi-chiller systems [Lee et al. 2011]. This study investigated the parameter setting issues of DE-based method, e.g., scaling factor and cross over factor. It was reported that the average performance of solutions was better than that of PSO-based method. Recently, some research has been conducted to study the usefulness of other evolutionary computing techniques. Dos Santos Ceolho and Mariani developed a modified firefly algorithm to minimize the energy consumption of multi-chiller system [Dos Santos Ceolho and Mariani 2013]. Two case studies were carried out. It was shown the proposed method could achieve better performance than other optimization methods, such as PSO and GA.

The above-mentioned studies mainly adopted linear regression or domain expertise to model the relationship between PLRs and the chiller energy consumption.

Since the intrinsic relationship may be highly nonlinear, the modeling process may not be accurate. The neuro-evolutionary method emerges to overcome such limitation. Chen, Chan, and Chan developed a neuro-evolutionary method to optimize the chiller loadings [Chen et al. 2014]. Rather than relying on linear regression, artificial neural networks were applied to modeling the highly nonlinear relationship between PLRs and chiller energy consumption. Then, PSO-based method was used to optimize the PLR of individual chillers. Comparison was made between the proposed method and linear regression with equal loading distribution method. It was shown that highly accurate results could be achieved with a faster convergence.

Multi-objective Optimization

Multi-objective optimization has been mainly used for HVAC systems. The objective function is usually formulated in such a manner that the total energy consumption is minimized while the indoor comfort is maintained.

Kusiak, Li, and Tang proposed a data-driven method for minimizing the energy consumption of a HVAC system, including chillers, pumps, fans, and reheat devices [Kusiak et al. 2010]. Eight DM algorithms were used to model the nonlinear relationship among room temperature, indoor relative humidity, CO_2 concentration, energy consumption of HVAC components, controllable parameters, and uncontrollable parameters. The set points of supply air temperature and static pressure in AHUs were optimized through PSO. It was shown that 7% of the total energy

consumption could be achieved.

Kusiak, Xu, and Tang proposed a DM-based method to optimize the HVAC system [Kusiak et al. 2011]. The predictive modeling was achieved by ensemble learning. A strength multi-objective PSO algorithm was used for optimization. Such algorithm is a combination of strength pareto evolutionary algorithm and conventional PSO algorithm. The results showed that better optimization solutions could be achieved than using conventional PSO method.

He, Zhang, and Kusiak presented a system-level optimization of HVAC system, with the aim of minimizing the energy consumption and the room temperature ramp rate [He et al. 2014]. Multi-layer perceptron models were developed to predict the energy consumption of AHUs, chillers, pumps, and fans. Two set points, i.e., the discharged air temperature set point and the supply air static pressure set point, were optimized to reduce the total energy consumption and ensure indoor comfort. Three optimization methods, i.e., evolutionary algorithm, PSO, and harmonic searching, were applied for control optimization. The research results showed that both harmonic search and PSO could be used for real-time optimization, and significant energy saving could be achieved.

West, Ward, and Wall adopted supervisory control and optimization techniques to optimize the HVAC systems of commercial buildings [West et al. 2014]. Predictive methods were developed to model the HVAC system power consumption, zone condition and thermal comfort. A weighted sum method was used to formulate the

multi-objective function, considering the energy consumption, greenhouse gas emissions, and occupant thermal comfort. Experiments showed that the significant energy saving and emission cuts could be achieved without sacrificing the indoor comfort.

2.4 DM Research and Applications for Building Design and Construction

2.4.1 Building Design

Building design has always been an important and complicated task. A good design leads to significant savings in both costs and energy use. Building design can be very challenging, as the number of design parameters is usually large and the actual performance are usually hard to quantify in advance. Building simulation software has been extensively used at the building design stage to predict building performance. It is capable of simulating the building operational performance based on different parameter settings and therefore, helps to make the decisions on optimal parameter settings. Typical design parameters include the building location, orientation, building envelope, heating, ventilation, and air conditioning (HVAC) system, lighting system, and etc.

There are two major challenges associated with the use of building simulation software. The first one is the large volume of simulation results. It is observed that

even a simple simulation could generate pages of data. Therefore, it can be very time-consuming to identify the most influential design parameters. Kim, Stumpf and Kim proposed a DM-based approach, which used the C4.5 decision tree algorithm, to identify the most significant design parameters [Kim et al. 2011]. A case study was carried out to investigate the design elements on four aspects, i.e., roof construction, wall construction, HVAC system, and the building orientation. In total, the research studied 127 options for roof construction, 88 for wall construction, 12 for HVAC, and 12 for building orientations. The results showed that HVAC has the largest impact on energy costs while the building orientation has the least impact. The proposed approach also evaluated the parameter importance in each sub-category. For instance, the insulation depth and the air space were identified as the most significant parameters in roof construction.

The second problem is about the optimization of parameter settings. Usually, there are a large number of building parameters and each parameter may have a number of possible values. Since the simulation is computationally expensive, it is infeasible to use building simulation software to try out every possible combination. To tackle this challenge, optimization techniques are normally used. Optimization aims to maximize or minimize the objective function by assigning values to a number of variables subject to predefined constraints. The use of optimization techniques helps to find optimal design parameters with a greatly reduced computation load. Through literature review, it is noted that optimization is the major technique used in

the building design stage. The following subsections summarize the representative works in the design of building envelope, building services systems and building-integrated renewable energy systems. More technical reviews on the optimization methods and algorithms can be found in [Evins 2013; Machairas et al. 2014].

Design of Building Envelope

The design of building envelope includes the selection of construction materials, the shape of the building, and the orientation and location. Leskovar and Premrov presented a shortcut to energy-efficient design of prefabricated timber-frame buildings [Leskovar and Premrov 2011]. A brute-force search was conducted based on the use the Passive House Planning Package (PHPP). The main aim was to determine the optimal proportion of glazing-to-wall ratio. Similarly, Goia, Haase, and Perino integrated simulation and optimization techniques to study the energy impact of building façade [Goia et al. 2013]. In their study, the performance of the façade was evaluated by the total energy consumption by heating, cooling, and lighting systems. One main design parameter to be optimized is the façade transparent ratio (i.e., window-to-wall ratio). The results showed that given state-of-the-art technologies, the transparent ratio has a low influence on the total primary energy demand. The minimum energy demand can always be achieved when the transparent ratio is between 35% and 45%. It is noted that since the optimization method is

brute-force, both the number and the resolution of the design parameters are restricted.

To overcome such restriction, meta-heuristic algorithms are commonly used. Tuhus-Dubrow and Krarti proposed a method to optimize building shape and envelope in terms of the building lifecycle costs [Tuhus-Dubrow and Krarti 2010]. The method integrated the genetic algorithm (GA) into the building simulation software *DOE-2*. The method enabled a global optimization with interactive effects considered. In total, 7 building shapes (i.e., L, T, H, U, Cross, Trapezoid) and 9 building envelope parameters (i.e., azimuth, aspect ratio, wall construction, ceiling insulation, thermal mass, infiltration, foundation insulation, window area, glazing type) were taken into account. The research results showed that across 5 different climates, the rectangular and trapezoidal shaped buildings consistently had the lowest lifecycle costs. Bichiou and Krarti integrated three optimization algorithms, i.e., genetic algorithm, particle swarm optimization (PSO) algorithm, and the sequential search algorithm, into the building simulation process to optimally select the building envelope features [Bichiou and Krarti 2011]. The building lifecycle cost was selected as the optimization target. The research results showed that GA and PSO required less computation time than the sequential search algorithm.

When there is more than one objective, multi-objective optimization methods are normally used. In general, these methods can be divided into two groups. The first one adopts the weighted-sum approach to transform multiple objectives into one

objective. In such a case, each objective function is normalized and assigned with a weight. The final objective function is a summed-up version of individual objective [Holst 2003].

The second type adopts the concept of Pareto optimal. A solution is Pareto optimal when there is no other solution available so that it will enhance one objective without deteriorating at least one of the others [Magnier and Haghighat 2010].

Design of Building Services Systems

Building services systems are closely linked to the indoor comforts and building lifecycle costs. Simulation software and optimization techniques have been widely used for the optimal design of the HVAC system and the lighting system. Staneuscu, Kajl, and Lamarche adopted a simulation-optimization method to design the HVAC system [Staneuscu et al. 2012]. The HVAC energy consumption was selected as the objective function and simulated by the software *DOE-2*. Evolutionary algorithm was then applied to optimize the design elements, i.e., grouping of the zones and the number of systems serving the building. The optimization results were encouraging, as significant energy saving could be achieved compared to the reference building and the existing building. Seo et al. adopted the multi-island genetic algorithm to optimize the HVAC system design [Seo et al. 2014]. The cooling and heating load of the HVAC system was first simulated by TRNSYS. Optimization algorithm was then applied to determine the type, number, and capacity of the HVAC equipment.

In terms of the lighting system, previous research mainly focused on two issues, i.e., daylighting system and artificial lighting system. For instance, Torres and Sakamoto investigated the availability of genetic algorithm for the optimization of daylighting systems [Torres and Sakamoto 2007]. The objective was to maximize the energy saving achieved by daylighting while preventing the discomfort glare. 21 parameters were selected for optimization, such as the window dimension, overhang depth, internal reflectance, and external reflectance. The proposed approach was reported to be effective. However, the thermal constraints were not integrated in the objective function. As a result, the optimized values for window area tended to be too large.

Design of Building-Integrated Renewable Energy Systems

The use of renewable energy helps buildings to achieve the goal of zero carbon emission or zero energy building. Many types of renewable energy can be generated on site. In this subsection, representative works related to the design of building-integrated renewable energy systems are reviewed, focusing on the design optimization of solar energy and geothermal energy systems. Solar energy system can be broadly categorized into two parts, i.e., solar thermal system and solar power system. Solar thermal system uses the solar radiation to heat up water or air. Solar power system transforms solar energy into electricity through photovoltaic panels. Hang et al. proposed a method, which consists of central composite design, regression

analysis, and multi-objective optimization, to optimize the design of solar absorption cooling and heating systems [Hang et al. 2013]. The multi-objective optimization considered the trade-off among energy, economic, and environmental aspects. The design parameters to be optimized included the slope and area of the solar collector, the volumes of the main storage tank and the hot storage rank, and etc. The proposed method was reported to be effective and can be used for other renewable energy systems, such as the wind or solar power systems.

Geothermal energy is another renewable energy source for buildings. The geothermal energy systems take advantage of the ground's relative constant temperature to provide heating or cooling energy to buildings. Kumar, et al. adopted the genetic algorithm to optimize the design parameters of an earth-to-air heat exchanger system [Kumar et al. 2008]. The design parameters, such as the mass flow rate, thermal conductivity of pipe, pipe length, and the radius of the tunnel, were optimized to maximize the cooling potential of the system.

2.4.2 Building Construction

In the past decades, more and more technologies, regulations and protocols have been developed to facilitate the building construction process. For instance, the newly developed technology, i.e., building information modeling (BIM) [Cerovsek 2011; Volk et al. 2014], has been used to store and simulate all the relevant data at the building construction stage. Such data has been used to facilitate the decision-making

at the construction stage (e.g., scheduling of construction tasks and estimation of activity duration.) [Kim et al., 2013]. DM techniques have been applied to extract useful knowledge from the construction data. The DM-related research and applications at the building construction stage can be broadly categorized into two groups, i.e., project management and occupational safety management. Representative research works are reviewed in the following subsections.

Project Management

Project management mainly deals with the management of project duration and project cost. An accurate prediction of these two variables helps the stakeholders to better prepare the bidding, schedule the project, and control the overall budget. In practice, such predictions are not easy to make. For instance, the project duration is affected by a number of factors, e.g., weather condition and resource availability. In terms of the project cost, the prediction can be of low accuracy due to the price fluctuations of labor, material, and equipment. Inaccurate predictions may lead to cost overruns, work interruptions, or even total project failure.

DM techniques have been applied to achieve better predictions. For instance, Son, Kim, and Kim proposed a hybrid method, which integrated principal component analysis (PCA) and support vector regression (SVR), to predict the construction cost [Son et al. 2012]. PCA was firstly applied to obtain a compact representation of the raw input data. A SVR model using radial basis function kernel was then developed to

predict the cost performance of commercial building projects. The model performance was proven to be optimal when compared with four other approaches, i.e., SVR, ANN, decision trees, and multiple linear regression. The prediction accuracy was reported to be accurate, with an MAPE of 10%.

The data collection during the construction stage can be very challenging due to the nature of construction process and the involvement of multiple parties. Consequently, the data to be mined are always insufficient and incomplete. To tackle these problems, Yu and Liu developed a method to mine scarce construction databases [Yu and Liu 2006]. The method was a hybridization of both symbolic reasoning and numeric reasoning. More specifically, the case-based reasoning method was integrated with two numeric reasoning methods, i.e., artificial neural network and neuro-fuzzy system. The proposed method was applied to handle five real-world construction problems. The proposed method was claimed to be effective in achieving higher mining accuracy while overcoming the limitation of data scarcity.

It is worth mentioning that compared to the data collected in the other stages in the building lifecycle, construction data are more diverse, containing both structured data and unstructured data. For instance, the data collected during the operation stage are typically stored in a structured way, using either spreadsheets or rational databases. By contrast, a large proportion of construction data are unstructured, such as document texts, project images, network-based project schedules. Knowledge extraction from the unstructured data is another hot topic in mining building

construction data. Soibelman, et al. reviewed relevant works on processing unstructured construction data [Soibelman et al. 2008]. A framework was proposed to manage and analyze text-based, web-based, image-based, and network-based construction databases. A preliminary study was carried out to demonstrate the usefulness of the framework. The framework was reported to be useful in enhancing the process of data management and knowledge discovery.

Occupational Safety Management

Building construction is a high-risk activity. Construction accidents typically lead to occupational injuries, fatalities, work interruptions, and other damages. DM techniques have been applied to identify the characteristic of occupational accident data. Based on the extracted knowledge, prevention strategies can be developed to reduce the chance of occupational accidents.

Classification and association rule mining are the most commonly used DM techniques in this area. Cheng et al. applied data mining techniques to examine the factors contributing to the occupational injury in the construction industry [Cheng et al. 2012]. A database with 1542 accident cases was analyzed using the classification and regression tree (CART) method. The derived occurrence rules successfully identified factors leading to different kinds of accidents. The findings could be used to improve the safety practices and protect construction workers from occasional or unexpected accidents.

Liao and Perng adopted association rule mining to explore the occupational accident reports [Liao and Perng 2008]. The associations between occupational injuries and fourteen contributing factors were discovered. The contributing factors included the individual factors, task factors, environmental factors, and management factors. It was reported that the proposed method could successfully identify the patterns of occupational injuries. The findings could be useful to establish effective inspection or prevention strategies. The authors also commented on the limitations of the method. As a large proportion of the obtained rules were useless, the post-mining of these rules was very time-consuming. It was recommended to develop a more advanced method to improve the efficiency in post-mining.

Cheng, Lin, and Leu used association rule mining to identify the cause-effect relationships in occupational accident data [Cheng et al. 2010]. Apriori algorithm was selected as the mining algorithm. A number of useful rules were successfully derived. It was found out that management and individual factors were the most significant ones contributing to the occupational accidents. The authors claimed that most accidents were preventable, as they were due the negligence of workers or management.

2.5 Summary

This chapter provides a comprehensive review on the DM-related research and

applications in building life cycle, with special emphasis on the building operation stage. Data mining (DM) has been successfully applied to a diversity of industries, including the retails, telecommunication, financial services, and even counterterrorism. DM techniques have found their strengths in three areas at the building operation stage, i.e., predictive modeling, fault detection and diagnosis, and control and optimization. Even though some encouraging results have been obtained, the potential of DM in the knowledge discovery in massive BAS data has not been fully exploited. Previous research relied heavily on domain knowledge across the whole mining process and mainly adopted supervised learning techniques. Consequently, the problem to be solved or the type of knowledge to be mined is usually predefined and only a small subset of BAS data was utilized. For instance, in the development of the prediction model for the chiller power consumption [Chang 2007], inputs to the model, e.g., the supply and return temperature of chilled water and condenser water were selected in advance, since domain expertise tells us that these variables are the most influential variables to chiller power consumption. The model developed may gave higher accuracy owing to the use of domain expertise and advanced supervised learning algorithms, the knowledge being discovered is usually limited.

In addition, the potential of unsupervised learning methods has not been fully discovered. Currently, only limited studies explored the usefulness of unsupervised learning methods in building operation, e.g., clustering [Khan et al 2013; Du et al. 2014] and association rule mining [Yu et al. 2012; Cabrera and Zareipour 2013]. The

main goal of DM is to discover potentially useful and previous unknown knowledge from massive data. Unsupervised learning methods are able to meet such goal by capturing the intrinsic structure, correlations, and associations from massive data, without explicitly defining the mining targets. Therefore, more effort should be made to investigate the usefulness of unsupervised learning methods. For instance, many studies of FDD are relied on the availability of training data. A training data consisting of both normal and faulty data is used to develop models for detection and diagnosis. However, such training data are very unlikely to be available in real practice. In view of this, unsupervised DM techniques may be more useful when tackling real-world problems.

Although DM technology brings valuable opportunity to effectively utilize massive building operational data, its applications in the building field still faces great challenges. DM itself cannot tell the value or the significance of the knowledge discovered. The knowledge discovered by DM is usually enormous and may be in various forms, such as predictive models, clusters, association rules, statistics. Meanwhile, advanced DM techniques are constantly emerging. It is not easy for building professionals to catch up the progress of DM technology. How to select the most suitable DM techniques and practically valuable knowledge are two big challenges. Based on the literature review, two general research directions are identified. The first is to keep exploring the vast amount of advanced DM analytics and identify their applications in building energy management. More specifically,

more research efforts should be paid on investigating unsupervised DM methods, as they are more flexible and capable of discovering potentially interesting and previously unknown knowledge. The other one is to develop generic DM-based frameworks for mining building operational data, including designing an overall workflow, developing detailed data preparation methods for a diversity of BAS variables, and devising post-mining methods for different types of knowledge representations (e.g., models, clusters and association rules).

CHAPTER 3 DEVELOPMENT OF DM-BASED ANALYTIC FRAMEWORK AND RESEARCH FACILITIES

This chapter introduces the DM-based analytic framework developed for the efficient and effective utilization of BAS data and the research facilities. Section 3.1 introduces the framework, which is deliberately designed by considering the unique characteristics of building data while ensuring the efficiency in the knowledge discovery process. Section 3.2 introduces the research facilities, which include the description of the real-world BAS data used for validation and the computation tools. A summary is provided in Section 3.3.

3.1 DM-based Analytic Framework

DM technologies are constantly evolving and the number of algorithms that are available to use is increasing rapidly. Meanwhile, it is realized that the system operations and the data collected from different buildings do share some similarities, such as the problem of poor data quality. Therefore, it is not wise to develop highly specific solution and investigate the usefulness of different approaches for each individual building. The establishment of a generic DM-based analytic framework

helps to standardize the knowledge discovery from building operational data and enhance the working efficiency of building professionals.

A typical knowledge discovery process usually includes five steps, i.e., data selection, preprocessing, transformation, data mining, and interpretation and evaluation [Han and Kamber 2011]. Based on the in-depth analysis of building data characteristics as well as considerations for practical applications, we develop a generic DM-based analytic framework for mining massive building operation data with five phases, i.e., data exploration, data partitioning, knowledge discovery, post-mining, and applications. The framework outline is shown in Figure 3.1 and the details of each phase are described in the following subsections.

3.1.1 Data Exploration

BAS data vary from building to building and the data quality can be poor due to the sensor errors and transmission malfunctions. Therefore, the first phase is data exploration, which aims to provide a general impression on the data behaviors and enhance the data quality. It is one of the most essential steps in a knowledge discovery process and it may take 80% of the total DM efforts [Zhang et al. 2003]. It consists of two tasks, i.e., data visualization and data preprocessing. Data visualization helps the users to visually gain preliminary understanding about the data. Visualization methods vary in their functionalities. For instance, box plots and

histograms are efficient to display the data distribution. Scatter plots provide a way to show correlations among variables. Run charts are useful to present time series data. Principle component analysis-based methods are useful for the visualization of high-dimensional data.

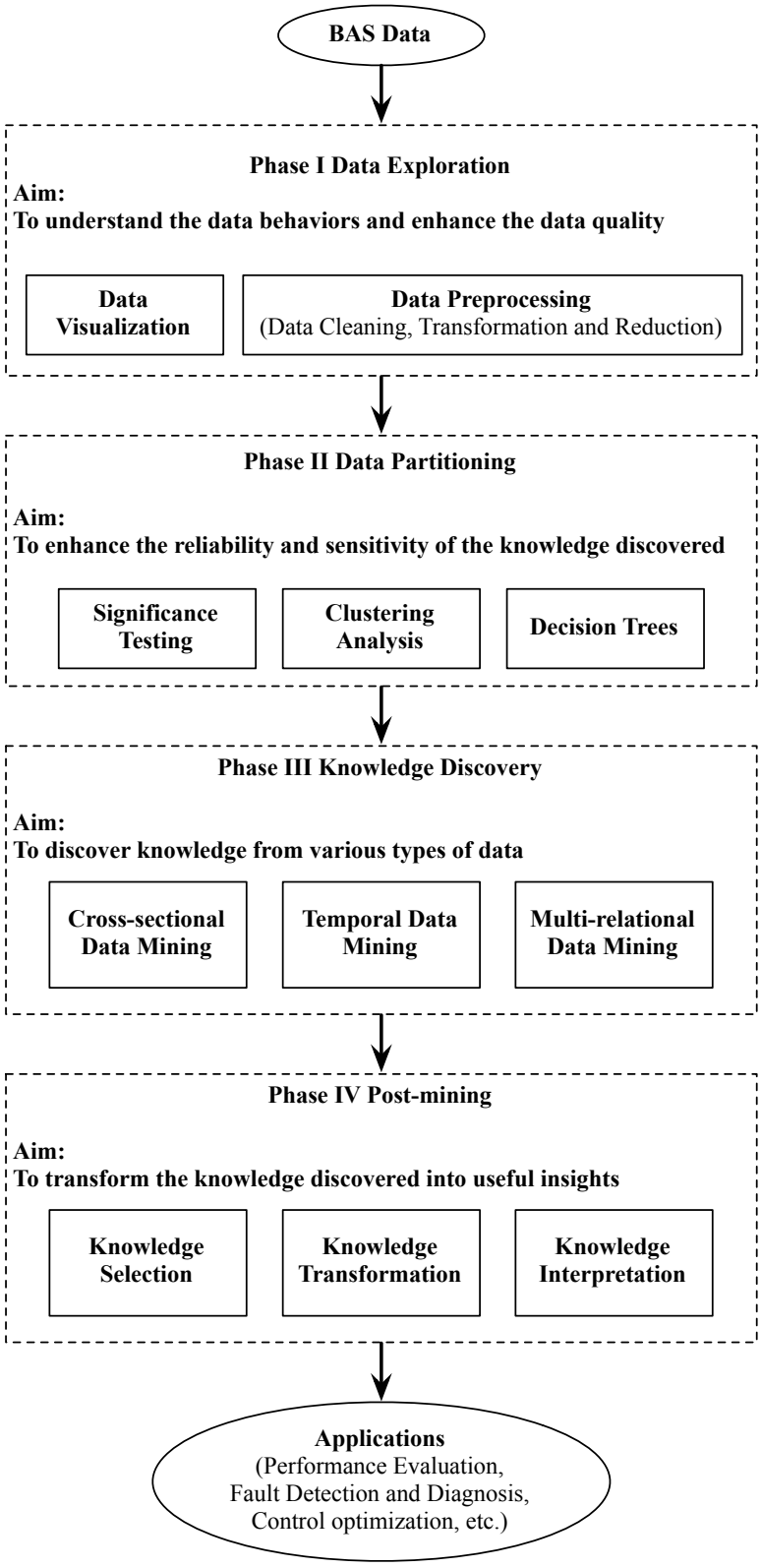


Figure 3.1 The DM-based analytic framework

The other key element in data exploration is data preprocessing. It typically involves three tasks, i.e., data cleaning, data transformation and data reduction. Data

cleaning aims to enhance the data quality by filling missing values, removing outliers and solve inconsistency in data. Missing values and outliers are two frequently encountered problems in BAS data, due to the sensor malfunctions or signal transmission errors. Moving average, imputation, and inference-based methods are frequently used in filling up the missing values. Outliers are those records that deviate from their true values. They can be solved using statistical methods, as well as unsupervised and supervised methods. Data inconsistency refers to the differences in the data scales or units, and unmatched records in different data sources. A popular solution to this problem is the data fusion schemes or physical redundancy [Huang et al. 2009].

Data transformation mainly consists of data scaling and data type transformation. Data scaling aims to normalize the data variables so that they appear equally important in data analysis as far as quantity concerned. Commonly used approaches include the max-min normalization, Z-score normalization, and decimal point normalization [Hastie et al. 2009]. Data type transformation is often needed when the data type is not compatible with the DM algorithms. For example, conventional association rule mining algorithms, such as the Apriori and frequent-pattern growth algorithms, can only deal with categorical data (e.g., High, Medium and Low), while the majority of the BAS data are numeric. Hence, it is necessary to transform numeric data into categorical data prior to the use of the conventional association rule mining algorithms. Popular methods for data type transformation include the equal-frequency

binning, equal-interval binning and entropy-based discretization [Hastie et al. 2009].

Data reduction aims to reduce the computation efficiency while maintaining the reliability and the generality of the knowledge discovered. BAS is usually stored in such a format that each row represents an observation sampled at a specific time instant and each column represents the values of a variable in all observations. Sampling techniques, such as random sampling and stratified sampling, are commonly used for the reduction of row number. The reduction of column number, or the selection of variables of interests and significance, can be done in three ways. The first one is to select the variables of interests based on domain knowledge. The second is to adopt data reconstruction methods, such as the principal component analysis in which the new low-dimensional variables are the linear combinations of the original high-dimensional data. The third is to use the heuristic methods, such as the step-wise forward selection and backward elimination methods, to select the most relevant to the problem concerned.

3.1.2 Data Partitioning

Data partitioning is necessary considering that most building services systems are highly dynamic and inter-correlated. The values of the variables and the relationships between variables may vary dramatically under different operation conditions. As a result, mining the entire BAS data simultaneously may result in

significant knowledge loss. Partitioning the BAS data into several subsets of unique patterns according to their intrinsic characteristics and then mining the individual subsets helps to efficiently discover more meaningful knowledge. Since the distance of the data in a subset is remarkably reduced, or the similarity of the data is greatly improved, the knowledge being discovered is more reliable. However, this kind of data partitioning should mainly rely on the data intrinsic characteristics and involves less domain knowledge to take the advantage of DM in discovering underlying knowledge. How to capture the data intrinsic characteristics is a critical issue and many methods can be adopted. In this research, three methods are explored to perform the data partitioning, i.e., significance testing, clustering analysis and decision tree methods. The details are shown in the later chapters.

3.1.3 Knowledge Discovery

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

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

More specifically, three methodologies are developed based on the generic DM-based analytic framework with the intention of discovering knowledge from cross-sectional data, temporal data and multi-relational data. The cross-sectional knowledge can be discovered by treating each observation in the BAS data as an independent event. Building operations are highly dynamic and intercorrelated. Therefore, a methodology for temporal knowledge discovery is proposed to discover the sequential patterns and temporal associations in BAS data. Currently, BAS data are usually recorded in a two-dimensional data table. It can be foreseen that BAS data structure will become more complex as more types of information are to be collected for the use of building management. In order to perform knowledge discovery from BAS data with potentially complex data structures (e.g., multi-relational data), a graph-based DM methodology is proposed. The details of these three methodologies

are described in the following chapters.

3.1.4 Post-mining

The post-mining phase aims to build a bridge between knowledge discovered in Phase 3 and practical applications, such as building performance assessment, fault diagnosis and optimization. The process can be very time-consuming, due to the large amount of knowledge discovered and the diversity of knowledge representations (e.g., rules, clusters, decision trees). Therefore, this research develops several methods to enhance the efficiency and effectiveness at the post-mining phase. The post-mining methods are specifically designed to couple with the DM algorithms used in the knowledge discovery phase. The methods are designed to perform three tasks, i.e., knowledge selection, knowledge transformation and knowledge interpretation.

Knowledge selection is to select potentially useful knowledge from the massive amount of knowledge discovered. Knowledge transformation aims to transform the knowledge into suitable formats for the ease of applications. Knowledge interpretation involves domain expertise to explain the knowledge discovered and thereby, converting the knowledge discovered into actionable measures for building operation management.

3.2 Research Facilities

3.2.1 Buildings and BAS Data

The BAS data retrieved from two buildings in Hong Kong are used to validate the performance of the DM-based analytic framework. The buildings and the BAS data are introduced in the following sections.

3.2.1.1 International Commerce Center (ICC)

ICC Building Description

Figure 3.2 presents the profile of the International Commerce Center (ICC). ICC is the highest building in Hong Kong. It stands 490m high with a total floor area of 321,000m^2. The building includes a 4-floor basement, a 6-floor block and a 112-floor tower. The basement has an area of 24,000m^2 and is mainly used as a parking area. The block building serves as a commercial center and the gross area is about 67,000m^2. Considering the tower building, the mechanical floors are located at the 6^{th}, 7^{th}, 42^{nd}, 78^{th} and 99^{th} floors, which are used to accommodate the chillers, cooling towers, water pumps, heat exchangers, PAU and AHU fans. The 8^{th}, 41^{st} and 77^{th} floors are used as refugee floors. The rest of the 9^{th} to 98^{th} floors are used as commercial offices, each with a length of 66m and a width of 65m. The 100^{th} to 118^{th} floors are used as a 6-star hotel.

Figure 3.2 The profile of International Commerce Center

BAS Data in ICC

The size of annual BAS data in ICC is around 30 gigabytes. The BAS data in ICC comes from two sources. The first mainly provides records of the sensor measurements and control signals of the central chilling system. The central chilling system consists of six identical high-voltage (10,000V) centrifugal chillers with a cooling capacity of 7,230 kW each. Each chiller is associated with a constant-speed primary chilled water pump and a constant-speed condenser water pump. The primary-secondary chilled water loop is used to transfer the cooling energy to the demand side. The heat dissipated from the chiller condensers is rejected by 11 evaporative cooling towers with a total design capacity of 51,709 kW. The data are collected at an interval of 1-min. In total, over 500 variables are recorded, such as the power consumption of chillers, pumps and cooling towers, and the temperatures, flow

rates of chilled and condenser water at different positions of the HVAC system water loop.

The second data source in ICC records the power consumption of a diversity of building components and services systems. Approximately 950 variables are collected at an interval of 15-minute. These variables can be aggregated to represent the power consumption of 5 main building services systems, i.e., the heating, ventilation, and air-conditioning (HVAC) system, normal power and lighting (NLTG), essential power and lighting (ELTG), vertical transportation system (VTS), and plumbing and drainage system (PD). The HVAC system in ICC consists of six subsystems, i.e., chillers, cooling towers, water pumps, primary air-handling units (PAU), air-handling units (AHU), and mechanical ventilation (MV). The vertical transportation system consist of the fireman's lifts, car parking lifts, escalators, office shuttle lifts and office services lifts. The NLTG includes the power consumption used for plug-ins and lighting. The power consumptions of the ELTG and PD systems are relatively small and constant during the building operation.

3.2.1.2 Zero Carbon Building (ZCB)

ZCB Building Description

Figure 3.3 presents an overview of the Zero Carbon Building (ZCB) in Hong Kong. ZCB has a total site area of 14,700m^2. The majority of the site is a landscaped area for public use. An eco-café and a small shop are located in the landscaped area.

The main component in the site is a 3-storey building with a footprint of 1,400m^2. It consists of an exhibition area, an eco-home, an eco-office, and a multi-purpose hall. Several passive design features have been integrated for energy saving, such as cross-ventilated layout, high performance glazing, light pipes and earth cooling tube. The active systems integrated include high-volume-low-speed fans, high temperature cooling system, intelligent lighting management, and absorption chiller, photovoltaic (PV) panels, and bio-diesel tri-generation systems. The estimated energy use of ZCB and the landscape area is around 116MWh and 15 MWh per year respectively. The major energy generation components are the biodiesel tri-generator and PV panels and their estimated energy outputs are 143 MWh and 87MWh per year respectively. More detailed information can be found in [CIC 2002].

Figure 3.3 The profile of Zero Carbon Building

BAS Data in ZCB

An intelligent BAS has been installed to monitor and control the building operational performance over various subsystems. Both the power consumption data of different building components and services systems, and the physical parameters of HVAC system are recorded with a collection interval of 1-hour. The power consumption data include 3 water-cooled chillers (WCC), 4 chilled water pumps (CHWP), 3 condenser water pumps (CDWP), 3 cooling towers (CT), 5 air-handling units (AHU) and 1 primary air-handling unit (PAU); the power consumption of outdoor landscape lighting (LandLight), the normal power and lighting consumption of the eco-office, basement area (Base), G/F common area (GF), multi-purpose room (MPR), mezzanine area (Mezz); the power generation the biodiesel tri-generator (BDG) and solar panels (SP) etc. The physical parameters of the HVAC system are also collected, including the temperatures and flow rates of chilled and condenser water at different positions of the HVAC water loops.

3.2.2 Computation Tools

A workstation has been equipped to perform the computing tasks involved in this research. The computer uses the Intel i7-3930K processor (12M Cache and 3.2GHz) and has a memory size of 16G.

The open-source computing software *R* is the primary mining tool for this research. Two *R* packages, i.e., "HighDimOut" and "TSMining", have been

developed to facilitate the mining tasks and their codes are presented in Appendix. Besides, the computing software *QuantMiner* [Salleb-Aouissi et al. 2007], SPMF [Fournier-Viger et al. 2014] and *ParSeMiS* [Worlein et al. 2011] are used to facilitate the tasks of temporal data mining and graph mining. The visualization tool *Gephi* and *Tableau* are used for data visualization.

3.3 Summary

This chapter introduces the DM-based analytic framework and research facilities. The framework serves as the generic solution for the knowledge discovery from BAS data. Three methodologies will be developed based on this framework to analyze different types of BAS data.

CHAPTER 4 DEVELOPMENT OF METHODOLOGY FOR CROSS-SECTIONAL KNOWLEDGE MINING AND ITS APPLICATIONS

The basic approach to the knowledge discovery from BAS data is to treat the data as cross-sectional data, in which each observation is treated as an independent event and the temporal dependency among observations is neglected. This chapter presents the methodology developed for mining cross-sectional knowledge in BAS data. Section 4.1 introduces the research methodology. The association rule mining is selected as the main knowledge discovery tool. Section 4.2 reports the typical operation patterns identified through data partitioning. Two possible approaches are developed based on the use of association rule mining methods. Sections 4.3 and 4.4 evaluated the usefulness of these two approaches using real-world BAS data. The last section 4.5 summarizes this chapter.

4.1 Research Methodology

The methodology is developed based on the generic DM-based analytic framework, as introduced in Chapter 3. It contains four main phases, i.e., data exploration, data partitioning, knowledge discovery and post-mining. The outline of

the research methodology is depicted in Figure 4.1. The details of the methods adopted at each phase are introduced as follows.

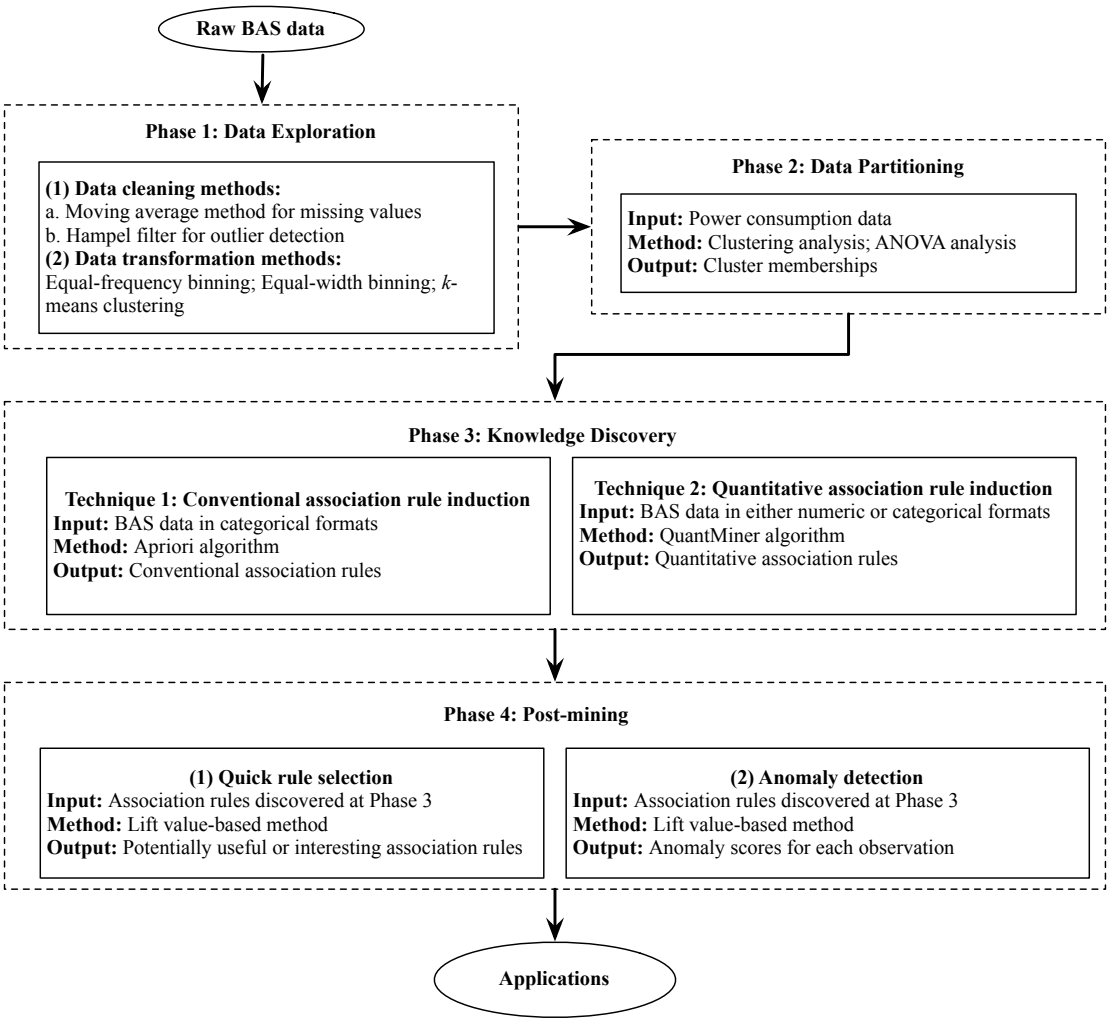


Figure 4.1 Research outline for cross-sectional knowledge discovery in BAS data

4.1.1 Data Exploration

Data Cleaning

The two main tasks involved in this methodology is data cleaning and data transformation. The data cleaning step handles the missing values and the outliers. The moving average method is one of the most widely adopted methods in filling up

missing values. It is easy to implement and have a fairly good performance when the duration of missing values is not long. It should be mentioned that if the missing values' duration is long, moving average methods cannot adequately simulate the dynamics in building operation. In such a case, it is recommended to either simply discard the data or use imputation-based methods to fill up the missing values.

Outliers are observations that are highly unlikely to occur based on the variation seen in the rest of the data. They can be classified into two types, i.e., points as outliers and subsequence as outliers [Gupta et al. 2014]. It is recommended that the data exploration phase only handles the first type of outliers in BAS data, as the identification of the second type of outliers may overlap with mining discords (i.e., infrequent sequential patterns) in the later process. The outlier detection methods can be grouped into three categories, i.e., prediction-based, profile-based, and deviant-based methods [Gupta et al. 2014]. The prediction-based methods detect outliers by comparing the actual measurements with their expected or predicted values from statistical analysis or machine learning algorithms. The profile-based methods use historical data to construct a normal profile, which is usually presented in the form of expected means and confidence intervals at different time. Each observation is compared with the normal profile to decide whether it is an outlier or not. The deviant-based methods identify outliers from a perspective of information theory. An observation is an outlier if removing it from the time series leads to a much more succinct representation of the original time series [Gupta et al. 2014].

In this methodology, the moving average method is used to impute the missing values with a short duration, i.e., less than 2-hour. Any missing values with a longer time duration are excluded from analysis. The Hampel filter is adopted to identify the first type of outliers, i.e., points as outliers. It is a nonlinear filter which shows high effectiveness in processing time series data [Pearson 2002]. For each observation, the Hampel filter calculates the median and the median absolute deviation (MAD) considering a moving window size of $2k+1$. k is the number of observations before and after the observation concerned. A parameter θ, which usually ranges from 0 to 5, is predefined to generate thresholds for outlierness evaluation, i.e., $Median \pm \theta \times MAD$. Any observation falls beyond the range of the thresholds is identified as an outlier and is replaced by the median. The smaller the parameter, the more aggressive the detection algorithm is and more observations will be identified as outliers. This study sets θ as 3, which is in accordance with the Ron Pearson's 3-sigma rule.

Data Transformation

Data transformation is carried out to cope with the use of conventional association rule mining algorithm at the phase of knowledge discovery. More specifically, numeric values in the BAS data are transformed into categorical. A number of methods are available. The equal-width method and equal-frequency method have been widely used due to their simplicity and reliability. The equal-width binning method divides the data into m intervals of equal size, while the

equal-frequency method divides the data into *m* groups which contains approximately same number of observations. Transformation results can greatly affect the mining performance. For instance, if the observation number in one category is too small, this category will be regarded as infrequent event. As a result, it may be very difficult to discover rules related to this category under a high support setting. There is no universally applicable guideline on how to select the optimal transformation method for a specific problem. It is recommended to examine the distribution of the numeric data first, and then integrate domain knowledge to select a suitable method for data transformation. For instance, the power consumptions of a two-speed fan can be easily categorized into three categories: *Low* (corresponding to zero power consumption when the fan stops), *Medium* (low speed) and *High* (high speed). Generally speaking, the more categories are used, the smaller the relative frequency of each category will be. Consequently, the support threshold should be set lower to cater for less frequent relationships when performing association rule mining.

In this research, the power consumption data and weather data are all numeric and they should be transformed into categorical data before mining association rules. Considering the climate conditions in Hong Kong, the outdoor air temperature is categorized into 6 levels with the interval of 5ºC from below 10ºC to above 30ºC, and the outdoor air relative humidity is categorized into 6 levels with the interval of 5% from below 70% to above 90%. The equal-frequency binning method, which results in an equal size of each category, is used to categorize all the power consumption data,

except for the power consumption data of PCHWP and CDWP. PCHWP and CDWP are constant speed pumps and their power consumptions will keep constant when the pumps run. Therefore, the power consumption data of PCHWP and CDWP are categorized according to the running pump number, for example, "2nd" means 2 pumps are running. The rest power consumption data are categorized into 3 categories using the equal-frequency binning methods, as they generally have a continuous distribution across their ranges. The three categories can be defined as *Low*, *Medium* and *High*.

4.1.2. Data Partitioning

Clustering analysis partitions the data into a number of clusters with the aim of maximizing the similarities of the observations in the same cluster while minimizing those between clusters. It is a natural fit for the task of data partitioning. The similarity can be measured by various methods, such as the Euclidean distance and the Manhattan distance. One potential obstacle in using clustering analysis for the task of data partitioning is that the results are not easily interpretable. It cannot directly output rules to gain insights into the data partitions obtained. Based on domain knowledge, it is realized that the building operations are closely linked to the time variables, e.g., Year, Month, Day and Hour. Therefore, a two-step approach is proposed to perform the data partitioning. Firstly, ANOVA is applied to analyze the

significance of time variables (i.e., month, date, hour, minute, day) to the target variables (e.g., aggregated power consumption). Then, clustering analysis is applied to find the optimal number of clusters which the original data can be partitioned into according to the significant variables, as well as to determine the cluster membership of each observation.

The clustering results can be evaluated by either the internal validation methods (e.g., Davies–Bouldin index, Silhouette index, and Dunn index) or the external validation methods (e.g., purity, F-measure, and normalized mutual information) [Tan et al. 2005]. In this study, the performance of five popular clustering analysis methods, i.e., k-means, partitioning around medoids (PAM), hierarchical clustering, entropy weighting k-means (EWKM) [Jing et al. 2007], and fuzzy c-means clustering, are compared. The parameters of these algorithms are fine-tuned using the Dunn index, which integrates the inter-cluster dissimilarity and cluster diameters to evaluate the clustering results. A larger Dunn index indicates a better clustering result.

4.1.3 Cross-sectional Knowledge Discovery

While the previous two phases prepare the data for mining, the knowledge discovery phase covers the actual mining process. A large number of DM techniques are available and new DM techniques are constantly emerging. The selection of DM techniques depends on the problems under consideration, data availability and the

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

Association rule mining (ARM) is a popular unsupervised DM technique and it has been adopted in retail, marketing, and health care [Salleb-Aouissi et al. 2007]. Compared with other forms of knowledge discovered by DM, interpretation of the association rules using domain knowledge is more convenient and utilization of the rules is more straightforward. Some efforts have been made on the application of ARM in the building field. Yu et al. [Yu et al. 2012] adopted the frequent-pattern growth algorithm to derive rules from the operational data of an air-conditioning system. The rules discovered were used to detect energy waste and component faults. Cabrera and Zareipour [Cabrera and Zareipour 2013] presented the application of

ARM in detecting lighting energy waste. The simulation results showed that up to 70% of energy use could be saved using the energy saving measures derived from the rules. There are two possible approaches to mining cross-sectional associations in BAS data. The first is to use the conventional ARM algorithms, such as the Apriori and FP-growth algorithms [Tan et al. 2005]. Such methods can only handle categorical data, such as "High", "Medium" and "Low". However, almost all BAS data, such as power, temperature, humidity, flow rate and pressure, are numeric. In practice, it is very difficult to determine the intervals for the categories of "High", "Medium" and "Low", since BAS variables generally present large varieties. The second is to adopt advanced association rule mining methods, i.e., the quantitative association rule mining (QARM), which are able to discover association rules in quantitative formats.

The usefulness of both approaches is evaluated and reported in the following sections. The background of these two types of association rule mining methods are given as below.

Conventional association rule mining

Association rule mining (ARM) is also an unsupervised learning process. It was firstly applied to perform the "market basket analysis", which aims to identify customer purchase behaviors. Later, ARM has been widely used to analyze large datasets in various fields, such as retail, bioinformatics and sociology [18]. The data to be mined by conventional association rule mining methods are usually required to

be categorical.

Let I be a non-empty item set, an association rule is a statement of the form $A \rightarrow B$, where $A, B \subset I$, and $A \cap B = \emptyset$. The set A is called the antecedent of the rule while the set B is called the consequent of the rule. Association rules are derived from a large number of observation sets (T), which is known as transaction sets in the DM field. Each variable or item in T belongs to I. Let $P(A)$ denote the probability that set A appears in the data set T and $P(A \text{ and } B)$ denote the probability that the sets A and B coincide in the data set T, the support, confidence and lift of an association rule are defined as Equations 4.1 to 4.3 respectively.

$$Support(A \rightarrow B) = P(A \text{ and } B) \tag{4.1}$$

$$Confidence(A \rightarrow B) = P(B|A) = \frac{P(A \text{ and } B)}{P(A)} \tag{4.2}$$

$$Lift(A \rightarrow B) = \frac{P(A \text{ and } B)}{P(A)P(B)} \tag{4.3}$$

ARM aims to find out all rules satisfying the user-specified minimum support or minimum confidence. Support of a rule is the joint probability of the antecedent and the consequent. Confidence is the conditional probability of the consequent, given the antecedent. Support and confidence are normally used to determine whether the rule is statistically significant or not. The support threshold can be defined with great flexibility. A higher support threshold tends to find rules that happen more frequently, and vice versa. A low support threshold will lead to a dramatic increase in the number of association rules obtained, and consequently the post-mining will be time-consuming. The confidence threshold should be maintained at a high level, e.g.,

above 85%, to ensure the association strength of the discovered rules.

Lift is a measure of the dependence and correlation between the antecedent and the consequent. It is usually used to evaluate the "interestingness" of an association rule. If the lift equals 1, it indicates that antecedent and the consequent are independent of each other, and hence, the discovered knowledge has little value. Lift larger than 1 indicates positive correlation, which means that the probability of the consequent is positively affected by the occurrence of the antecedent. In contrast, lift smaller than 1 indicates negative correlation. Generally speaking, the larger the lift value deviates from 1, the more interesting the rule is.

Quantitative association rule mining (QARM)

The rule format for quantitative association rules is as follows: $\{A \in [a_1, a_2]\} \rightarrow \{B \in [b_1, b_2]\}$, where A and B are numeric variables, and a_1, a_2, b_1, b_2 specify the intervals for each numeric variable. $\{A \rightarrow B\}$ is called the rule pattern. Similarly to conventional association rule mining methods, quantitative association rules are derived by defining two parameters, i.e., the minimum thresholds of support and confidence. Only those rules meet the thresholds are derived and considered to be meaningful.

The QARM algorithm adopted in this research is called the QuantMiner [Salleb-Aouissi et al. 2007; Salleb-Aouissi et al. 2013]. The intervals (a_1, a_2) and (b_1, b_2) are determined by compromising the gain of an association rule and the length of the intervals. The gain of an association rule is defined by Equation (4.4), where

MinConf is the predefined minimum confidence threshold. The fitness function, which takes into account both the gain of the association rule and the length of the intervals, is defined by Equation (4.5). Genetic algorithm is used to maximize the fitness function. Rules with large gains and small intervals are preferred [Salleb-Aouissi et al. 2007; Salleb-Aouissi et al. 2013].

$$Gain(A \rightarrow B) = Support(A \cap B) - MinConf \times Support(A) \qquad (4.4)$$

$$Fitness(A \rightarrow B) = Gain(A \rightarrow B) \times \coprod_{A_i \in A_{num}} \left[1 - \frac{size(I_{A_i})}{size(A_i)}\right]^2 \qquad (4.5)$$

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

4.1.4 Post-Mining

The post-mining method described in this section is used for mining quantitative association rules. It handles two specific challenges, i.e., rule selection and rule utilization. The details are shown as below.

Rule selection

As mentioned above, the support and confidence are used to evaluate rules, and only those rules with the support and confidence meeting the predefined thresholds are considered. However, hundreds of rules may still be obtained, although the thresholds of the support and confidence are conservatively set. Selecting potentially

useful rules by individual inspection is extremely time-consuming. It is noticed that the lift is helpful in selecting potentially useful rules. The larger the lift value deviates from 1, the more interesting the rule is. A novel rule selection approached based on the lift is proposed for fast selection of potentially useful rules.

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

Rule utilization

The knowledge discovered by DM can be used for various purposes, including prediction, diagnosis, and optimization. A method for utilizing the association rules for diagnosing abnormality in operation is proposed. All the rules obtained from mining one subset are utilized to build the knowledge base for the corresponding operation condition. Each new observation is examined against the rules in the corresponding knowledge base. A rule is violated if the observation meets the antecedent but fails to meet the consequent. Since the lift value indicates the dependence strength between the antecedent and the consequent, the rules with larger lift values are more significant than those with smaller lift values. As a result, if an observation violates a rule with larger lift, the violation is more serious. Accordingly, an abnormality degree (AD) of an observation is proposed as shown in Equation (4.6), which measures the seriousness of the violation against all rules in the corresponding knowledge base. In Equation (4.6), 1 is subtracted from the lift values, as a lift value of 1 indicates independence between the antecedent and the consequent. A lift value smaller than 1 means the probability of occurrence of the consequent is low when the probability of occurrence of the antecedent is large. Therefore, if a new observation violates a rule with a lift value smaller than 1, it is actually normal, rather than abnormal, and the AD should be decreased.

$$AD = \sum_{i=1}^{n} (lift_i - 1) \qquad (4.6)$$

Where n is the number of rules being violated, and $lift_i$ is the lift value of the ith

rule being violated.

4.2 Identification of Typical Building Operation Patterns

Phase 2 performs the data partitioning with the aim of identifying typical building operation patterns. Building operation is mainly influenced by climate conditions and occupancy level. Moreover, people are very much concerned about building energy efficiency and indoor environment quality (IEQ). Identification of building operation patterns related to energy consumption and IEQ can help to explore means to enhance them. IEQ can be assessed by monitoring the concentrations of the indoor CO_2 and other typical pollutants, indoor illuminance and noise levels, and etc. However, such measurements are usually not available in today's BASs, including the BAS of ICC. The power consumptions of various components are well recorded in ICC. Therefore, this research focuses on typical power consumption patterns.

As described in Section 4.1.2, the identification process is undertaken in two steps. Firstly, ANOVA is applied to analyze the significance of time variables (i.e., month, date, hour, minute, day) to the aggregated power consumption. Then, clustering analysis is used to find the optimal number of clusters which the original data can be partitioned into according to the significant variables, as well as to determine the cluster membership of each observation.

The level of Type *I* error, α, is defined as 1% to make the results more stringent. The ANOVA results were shown in Table 4.1. It indicates that only three time variables, i.e., month, day, and hour, result in a probability smaller than the specified Type *I* error. Therefore, these three variables have significant effects on the aggregated building power consumption. Five clustering analysis methods (i.e., k-means, hierarchical clustering, PAM, fuzzy c-means, and EWKM) are adopted to partition the large BAS data. The clustering analysis is performed in the sequence of "month", "day", and "hour" to avoid the conflict of cluster memberships. To determine the cluster membership in terms of the variable "month", all the power consumptions of the 12 sub-systems are scaled using max-min normalization. Then, the mean and standard deviations of the power consumption of each sub-system are calculated for each month, resulting in a feature data set with 24 variables. Clustering analysis is performed based on the feature data set. When determining the cluster membership in terms of the variable "day", the original observations in the months which are grouped in the same cluster are analyzed together. Features are then calculated for each day (i.e. Monday to Sunday) and clustering analysis is applied to find the cluster membership. Similar approach is adopted in determining the cluster membership in terms of "hour".

Table 4.1 ANOVA testing results

Variable	DOF	Sum of squares	Mean sum of squares	F-test statistics	Probability

Month	11	34,233,661	3,112,151	759.46	< 1%
Date	30	1,361,300	453,800	0.37	78%
Hour	23	220,948,230	9,606,445	2,344.25	< 1%
Minute	3	2,401	800	0.20	90%
Day	6	89,532,444	14,922,074	3,641.42	< 1%

Figures 4.2 and 4.3 illustrate the clustering results in terms of "hour". Figure 4.2 presents the Dunn indices of different clustering algorithms and cluster numbers. The maximum Dunn index can be obtained when EWKM is used and the cluster number is 2. Therefore, such combination is selected to perform the clustering analysis. Two parameters of EWKM, i.e., the weight distribution parameter (λ) and the convergence threshold (δ), are determined using the Dunn index and they are 0.15 and 0.0001, respectively. Figure 4.3 illustrates the cluster membership in terms of the time variable "hour". It can be found that the majority of observations collected during 8:00 to 20:00 are grouped in the 1st cluster and the rest observations are grouped in the 2nd cluster. In Hong Kong, 8:00 to 20:00 are normally the office hours and the other hours are non-office hours. The power consumptions during office hours and non-office hours are very different due to the different operation conditions, particularly occupancy levels. Therefore, the clustering result is consistent with the domain knowledge.

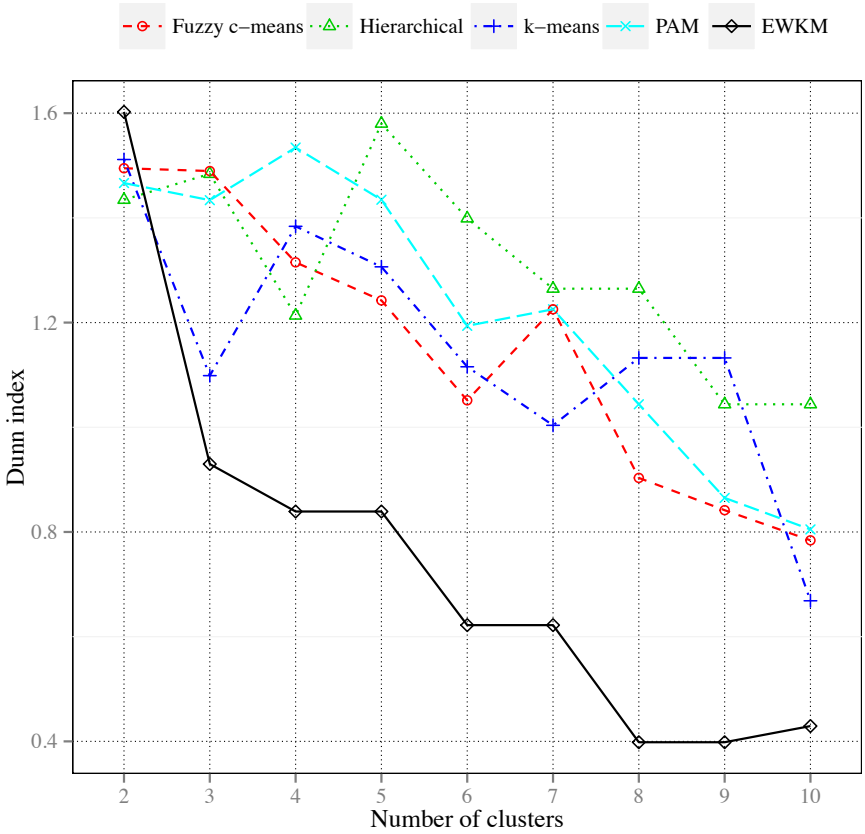


Figure 4.2 Comparison of clustering algorithms for clustering in terms of "hour"

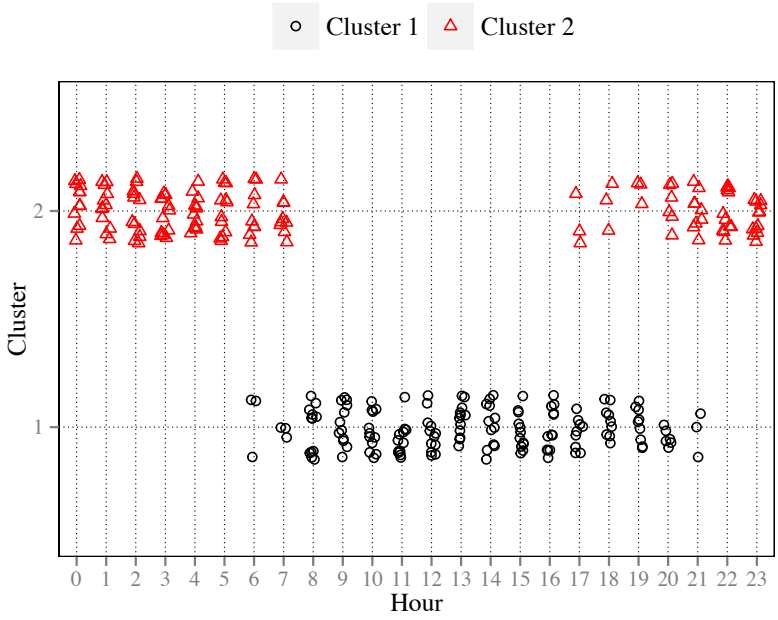


Figure 4.3 Cluster membership in terms of “hour”

Table 4.2 summarizes the overall clustering results. The optimal cluster number is 2 for all cases. EWKM is chosen as the clustering algorithm in terms of "month" and "hour", while the hierarchical method is chosen for "day". When the data is grouped in terms of "month", data collected from June to October are grouped in one cluster, and the data from the other months are grouped in the second cluster. June to October are normally the hot season in Hong Kong with the higher outdoor temperature and relative humidity. Cooling demand in hot season is very large. By contrast, the other months are cool season and cooling demand is relatively low. The two clusters in terms of "day" are corresponding to weekdays (i.e. Monday to Friday) and weekends (i.e. Saturday and Sunday), respectively. As a result, the original large BAS data are partitioned into eight clusters or subsets, and each cluster is defined by a combination of the three time variables as shown in Table 4.3. The clustering results are reasonable, since each cluster has its unique power consumption pattern considering the climate conditions and occupancy levels.

Table 4.2 Summary of the clustering results

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

Table 4.3 Summary of the eight clusters

Cluster	Month type	Day type	Hour type
1	Hot season	Weekdays	Office hours

2	Hot season	Weekdays	Non-office hours
3	Hot season	Weekends	Office hours
4	Hot season	Weekends	Non-office hours
5	Cool season	Weekdays	Office hours
6	Cool season	Weekdays	Non-office hours
7	Cool season	Weekends	Office hours
8	Cool season	Weekends	Non-office hours

4.3 Discovery and Applications of Qualitative Associations

As described above, eight typical building operation patterns were identified. The transformed data sets, i.e. with numeric values transformed into categorical values, were divided and mined accordingly. The Apriori algorithm was selected to discover associations in BAS data.

Two key parameters, i.e., minimum support and confidence, should be determined to carry out the ARM. In this study, the minimum support is set relatively low, i.e., 0.1, to capture associations between infrequent events. By contrast, the minimum confidence threshold is set to be relatively high, i.e., 0.85, to ensure the reliability of obtained rules. Considering that the interpretability of discovered rules decreases with the increase in item number, the minimum and maximum item number (i.e. the total number of antecedents and consequents) in a rule was set to be 2 and 5, respectively. Redundant rules were removed by comparing their lift values. For instance, assuming that Rule A and Rule B have the same consequent, and Rule A's

antecedent is a superset of Rule B's. If Rule A has the same or a lower lift value, Rule A is redundant and removed.

In total, 457 assocaition rules were derived. Most of rules can be easily obtained from domain knowledge and hence be ignored in this study. For example, Rule 1 in Table 4.4 describes that, if the outdoor temperature on Saturday is between 15 ºC and 20 ºC, the chiller power consumption is "Low". This can be easily understood, as a low outdoor temperature and a low occupancy level on Sunday always lead to a small cooling load and hence a low chiller power consumption. Rule 2 states that, if the power consumption of the primary air handling units is "High", the power consumption of lifts is "High". This rule can also be easily interpreted, as both the power consumption of the primary air handling units and lifts are closely related to occupancy level. A higher power consumption of the primary air handling units normally indicates a higher occupancy level and hence, more people need to use the lifts for vertical transportation. Four representative rules, which are either against

common experience or of particular value, are analyzed in detail.

Table 4.4 Examples of association rules discovered

4.3.1 Detection of Deficit Flow

Rules 3 and 4 are interesting because they disobey one simple design principle.

No.	Antecedent	Consequent	Supp.	Conf.	Lift	Cluster
1	Out.T=(15,20)	Pwr.Chiller=Low	0.25	0.88	2.10	1
2	Pwr.PAU=High	Pwr.Lift=High	0.35	0.86	1.76	1
3	Pwr.PCHWP=4th	Pwr.CDWP=3rd	0.27	0.99	2.73	1
4	Pwr.PCHWP=3rd	Pwr.CDWP=2nd	0.32	0.88	1.78	3
5	Pwr.PCHWP=4th	Pwr.SCHWP=High	0.24	0.89	2.83	1

In the design, each chiller is associated with one constant-speed primary chilled water pumps (PCHWP) and one constant-speed condenser water pumps (CDWP). The running numbers of the PCHWP and the CDWP should be the same as the number of the chillers in operation, so called one-to-one operation strategy. Rule 3 indicates that in Cluster 1 (i.e., Hot Seasons, Weekdays, Office hours), if the PCHWP power consumption is at the 4^{th} level (i.e., 4 PCHWP are running), the CDWP power consumption is at the 3^{rd} level (i.e., 3 CDWPs are running). Rule 4 states a similar phenomena in Cluster 3 (i.e., Hot Seasons, Weekends, Office hours). If the PCHWP power consumption is at the 3^{rd} level, the CDWP power consumption is at the 2^{nd} level. There is always one more PCHWP in operation, which may cause significant energy waste. Therefore, the operational strategy of PCHWPs should be investigated.

Figure 4.4 shows the relative frequency when the same number of PCHWPs and

CDWPs are in operation in each month. The relative frequency is the number of events concerned divided by the total number of observations. It can be found that from May, the numbers of PCHWPs and CDWPs in operation are different for more than 90% of the time. After checking with the operation staff, the reason was found. To prevent deficit flow, one extra PCHWP was started to compensate the flow rate in the primary loop. This deficit flow prevention strategy was implemented occasionally before May; however, it has been consistently used starting from May.

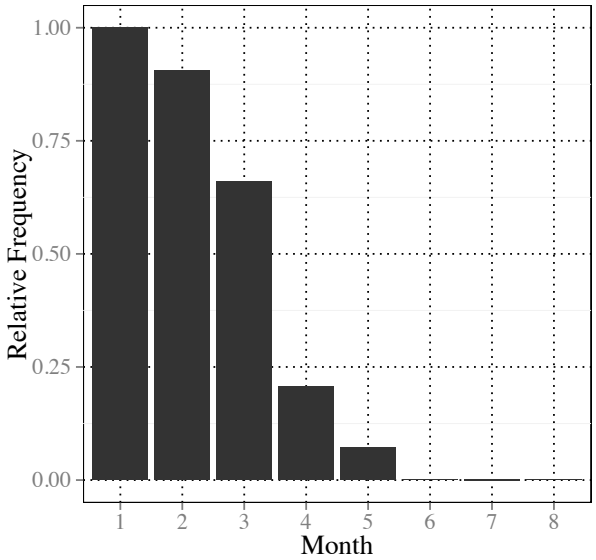


Figure 4.4 Relative frequency of one-to-one operation conditions

Deficit flow is a commonly encountered problem in the primary-secondary chilled water systems with decoupled bypass line. It normally takes place when the required flow rate of the secondary loop exceeds that provide by the primary loop. Severe operational problems can be caused, such as high supply chilled water temperature, over-supplied chilled water, and increased energy consumption of secondary pumps [Gao et al. 2011].

It is obvious that the current operation strategy is not energy-efficient due to the operation of one extra PCHWP. However, it seems necessary to operate one more PCHWP to prevent deficit flow. The question is whether the operation strategy can effectively prevent deficit flow, or the cost of energy is worthwhile or not. To answer this question, recursive partitioning was applied to evaluate the effectiveness of such a strategy. All the observation sets under the condition that the number of running PCHWPs equals the number of running CDWPs plus 1 were extracted for further analysis. 13004 observation sets in total were obtained. The flow rate in the bypass line was selected as the indicator of deficit flow. If it is negative, deficit flow occurs, and vice versa. A tree model was built using recursive partitioning. The model output is either "Deficit Flow" or "Normal Condition". The confidence level was set to be 95% in determining the splitting variable. To optimize the configuration of tree model, two parameters, i.e., the minimum number of observations in a node to perform splitting and the minimum number of observation in a terminal node, were specified. These two parameters were determined by cross-validation using the classification purity as evaluation criteria. As a result, these two parameters were set as 3000 and 1000 respectively.

The developed conditional inference tree is shown in Figure 4.5. Each terminal node shows the proportion of the classified items, "D" for deficit flow and "N" for normal condition. The rated power of each PCHWP and CDWP are 126kW and 202kW, respectively. The first two terminal nodes, Node 3 and 4, indicate that deficit

flow still occur frequently when the number of running PCHWP is 2. More specifically, if the outdoor temperature is relatively high, i.e., higher than 22.95ºC, deficit flow always occur. By contrast, if the outdoor temperature is relatively low, i.e., lower than 22.95ºC, the chance of deficit flow decreases. Therefore, when only 1 chiller is in operation, deficit flow cannot be prevented effectively by running one extra PCHWP. In addition, Node 10 also indicates that the operation strategy cannot effectively prevent deficit flow in the corresponding situation. It shows that when the number of running PCHWP is 3 and the chiller power consumption is larger than 1861.7kW, which is between the capacities of one and two chillers, deficit flow occurred in 60% of the operation time.

The other three terminal nodes, Node 8, 9 and 11, indicate a good performance as deficit flow can be prevented effectively. It is noticed that variable "Month" is selected as the splitting variable for Node 7. It is observed that before May, when the running number of chiller is 2 and the running number of PCHWPs is 3, no deficit flow occurs. By contrast, starting from May, under the same condition, deficit flow may occur with around 15% probability. The potential affecting factors can be climate, system set points, and etc. Node 11 shows that when the running number of PCHWPs is larger than 4, no deficit flow occurs. It can be concluded that the current operation strategy is effective to prevent deficit flow when 3 or more chillers are in operation.

To sum up, the recursive partitioning model reveals that the current operation strategy for preventing deficit flow is not effective, particularly when only one chiller

is in operation, or low cooling load condition. It is recommended to develop different control strategies for low cooling load condition to effectively overcome the problem of deficit flow and save energy.

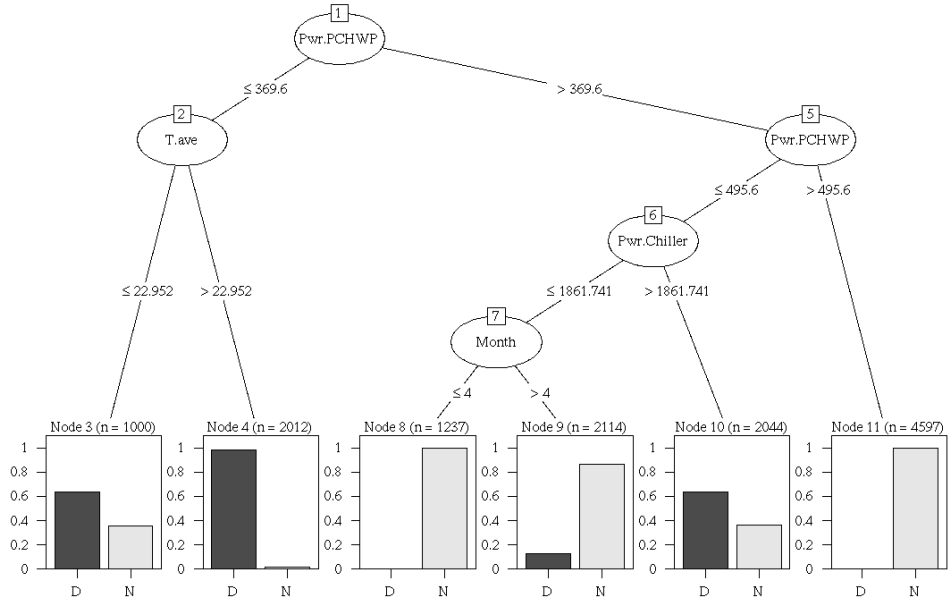


Figure 4.5 Developed conditional inference tree

4.3.2 Detection of Abnormal Operations

Rule 5 in Table 4.4 is derived from Cluster 1 (i.e., Hot seasons, Weekdays, Office hours). It says that if the PCHWP power consumption is at the 4th level, the secondary chilled water pump (SCHWP) power consumption is “High”. Rule 5 shows a reasonable relationship between the energy consumption of primary pumps and secondary pumps. As the cooling load increases, the required secondary chilled water flow rate increases and the power consumption of SCHWPs increase, too. When the cooling load increases significantly, one more chillers and hence one more PCHWP

are started. Therefore, more PCHWP in operation means a greater cooling load and hence more power consumption of SCHWPs. Although this rule can be easily understood with domain knowledge, the quantitative description of the rule is not that straightforward. ARM provides an applicable rule for detecting abnormal operation of the primary and secondary pumps.

Using this rule to examine the raw data, it was found that the weekday's data sets have 438 abnormal observations. It was also found that the majority of these abnormal observations are sparsely distributed on different days, resulting less than 5 (i.e., 75 minutes) continuous abnormal observations for one specific day. Since HVAC system may experience transient situations during the On-Off control of the major components like chillers and pumps, these sparse abnormal observations can be ignored. However, if a large number of abnormal observations occur continuously as described in the example below, it is reasonable to believe that the operation presents some problems.

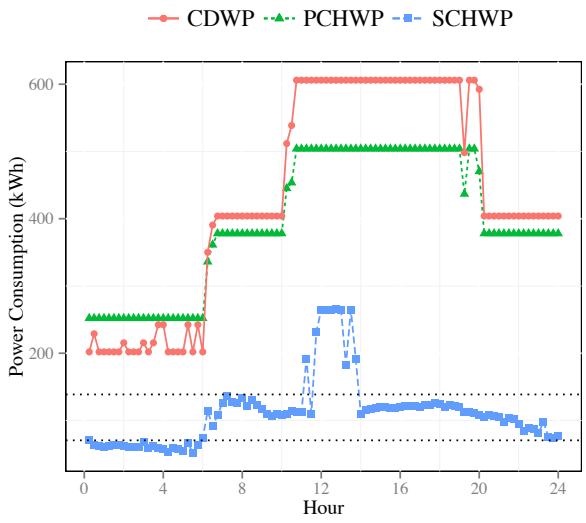


Figure 4.6 Abnormal operation condition

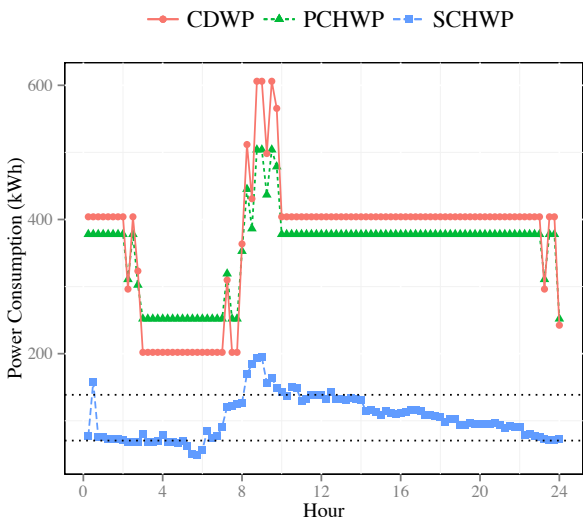


Figure 4.7 Normal operation condition

Figure 4.6 shows the abnormal primary-secondary pump operation in one weekday founded in the raw data sets. Starting from 11:00, the SCHWP power consumption undergoes a rapid increase and its running condition is changed from "Medium" to "High". At the same time, the PCHWP power consumption rises to around 500 kW, which corresponds to the power consumption of 4 PCHWPs (i.e. 4×126kW=504kW). The CDWP power consumption also rises to around 600kW, which corresponds to the operation of 3 CDWPs (i.e., 3×202kW=606kW). Three hours later, i.e., at 14:00, the SCHWP power consumption drops back to the "Medium" level and never reaches the "High" level again during the rest of the day. Nevertheless, no action was taken for PCHWPs and CDWPs, as they keep on running with high intensity until 20:00. The operation from 14:00 to 20:00 does not satisfy Rule 5 and can be diagnosed as abnormal operation. In this case, it wastes energy in the 6 hours.

The extra energy cost is around 2,000 kWh (i.e., (126+202)×6=1,968kWh) on that day.

By contrast, Figure 4.7 shows the normal operation. The SCHWP power consumption reaches the "High" level at 08:00 and it drops back to the "Medium" level around 2 hours later. Corresponding to these changes, it is observed that one more PCHWP and CDWP are switched on at around 08:00 and switched off at 10:00.

4.4 Discovery and Applications of Quantitative Associations

QuantMiner is selected to mine the eight data subsets separately. The minimum thresholds of the support and confidence are also set as 0.1 and 0.85. Although QuantMiner is capable of mining association rules with multiple variables in both antecedents and consequents, this research only focuses on the association rules with only one variable in the antecedent and consequent respectively for easy interpretation.

Each of the eight subsets generates 534 rules and 4,272 rules are obtained in total. The post-mining method described in section 4.1.4 is adopted for rule selection, interpretation, and utilization. The SD-Lifts of similar rule patterns obtained from mining different subsets are calculated. It is found that the majority of the rules with the same rule patterns have a SD-Lift smaller than 0.2. The rule patterns having a large SD-Lift have been selected for further analysis. An example is presented in the

following section.

4.4.1 Identification of Change in Building Operation Strategies

One rule pattern $\{WCC \rightarrow PAU\}$ draws special attention as it has a large SD-Lift of 0.3. $\{WCC \rightarrow PAU\}$ describe the associations between the chiller power consumption (WCC) and the PAU fan power consumption (PAU). The details of the rules with this rule pattern are shown in Table 4.5. The clusters are numbered in accordance to Table 4.4. The rules obtained from mining Cluster 1 and 8 are interpreted with domain knowledge here. Cluster 1 is corresponding to "Hot season", "Weekdays" and "Office hours", which means hot climate and high occupancy level. Cluster 8 is corresponding to "Cool season", "Weekends" and "Non-office hours", which means cool climate and low occupancy level. According to domain knowledge, the demand for cooling is higher under hot climate, and the demand for outdoor air ventilation is higher for high occupancy level. Therefore, both WCC and PAU of Cluster 1 should be higher than those in Cluster 8, which can be seen from the intervals of WCC and PAU of Cluster 1 and Cluster 8 in Table 4.5. Meanwhile, if the "day type" and "hour type" are the same which means the occupancy level are similar, the WCC intervals in "Hot season" should be higher than those of "Cool season". Rules obtained from Cluster 1 and Cluster 5 also support this argument. However, it

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

Table 4.5 Summary of the rule pattern {WCC → PAU}

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

A decision tree is developed to explore the underlying relationship among time variables and the PAU fan power consumption. The PAU fan power consumption is the output, while the "month", "day" and "hour" are selected as inputs. For easy

interpretation, the tree depth is limited to 2, which means the remotest terminal node can be reached from the root node through 2 splits. As shown in Figure 4.8, four terminal nodes (i.e., Node 3, 4, 6, and 7) are derived to represent the four levels of PAU power consumption. The associated boxplots show the distribution of the PAU power consumption at each terminal node. The algorithm selects the "hour" as the splitting variables at the root node (i.e., Node 1). It automatically divides the "hour" into two groups, i.e., non-office hours {0, 1, 2, 3, 4, 5, 6, 7, 21, 22, 23} and office hours {8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20}, which is in accordance with the cluster membership discovered by clustering analysis. Similarly, Node 5 selects the "day" for splitting and the results are the same as that obtained from the clustering analysis. Node 2 uses the "month" for splitting; however, the grouping of months (June to December in one group, and January to May in the other group) is different from the clustering results (June to October in one cluster, and the other months in the other cluster).

The right side of the tree states that when the observations are measured during the office hours, the PAU power consumption is closely related to the "day type". It is observed that the PAU power consumption in weekdays is significantly higher than that in weekends. This is reasonable because people normally don't work in weekends which results a large drop in the occupancy level. The left side of the tree states that when the observations are recorded during the non-office hours, the PAU power consumption is closely related to the "month". It is noted that the PAU power

consumption during the first five months (i.e., Jan to May) is higher than that during June to December, which cannot be explained by domain knowledge.

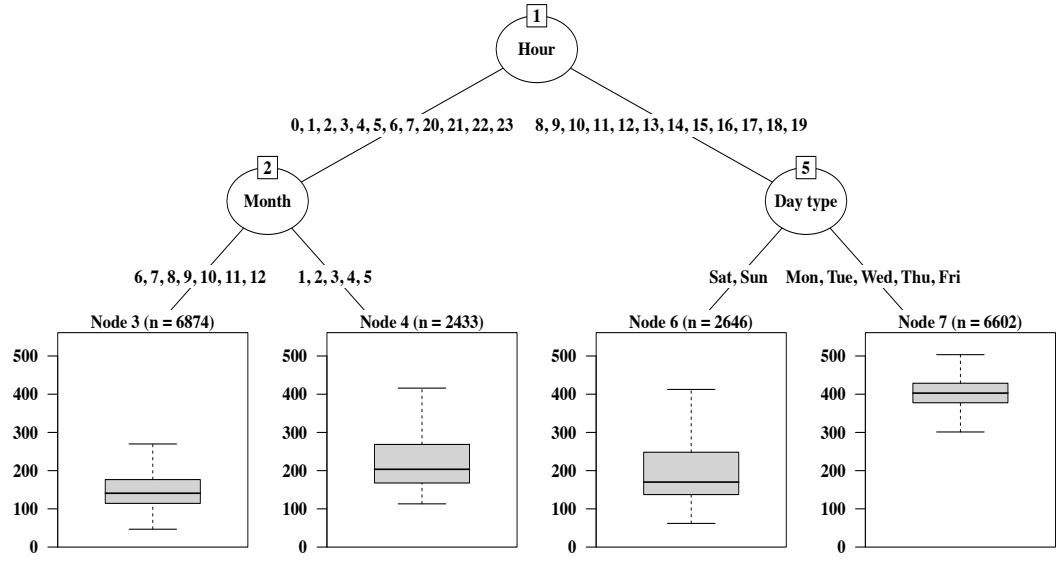


Figure 4.8 Decision tree for PAU power consumption

After consulting the operation staff, it is found out that the PAU operation strategy did change in June 2013. Before the change, the PAU fan speed was controlled at three levels, i.e., 0 L/s, 960 L/s, and 1200 L/s. If the CO_2 concentration is below 800 ppm, 960 L/s is used; otherwise, 1200 L/s is used. Starting from June 2013, the demand-controlled ventilation (DCV) strategy is implemented. Under this strategy, the fresh air flow rate is continuously controlled between 850 L/s and 1200 L/s to maintain indoor CO_2 concentration at its set-point. That's why the tree model adopts the "month" as the splitting variable at Node 2, which also indicates that the DCV strategy results in more energy saving during non-office hours. This is reasonable because that the occupancy level during non-office hours is quite low and the demand

for outdoor air ventilation is also low. The energy saving during the office hours is not that obvious because the occupancy level is quite stable throughout the year.

4.4.2 Identification of Atypical Building Operations

A number of continuous observations in Cluster 5 (i.e., Cool season, Weekdays, Office hour) are found to have high abnormality degrees (ADs). The examples of the rules being violated are summarized in Table 4.6. "Temp_rtn_ch" and "Temp_sup_ch" refer to the return and the supply chilled water temperature respectively. The first two rules state that the NLTG power consumption has associations with the WCC power consumption and the return chilled water temperature. The lower limits of NLTG are approximately 500 kW. However, the actual NLTG measurements of the observations with high ADs are around 430 kW. The 3rd and the 4th rules describe the relationship among PAU fan power consumption, the supply chilled water temperature, and WCC. The lower limits of PAU are around 300 kW, while the PAU fan power consumptions in the observations with high ADs are around 250 kW. The last two rules describe the associations among VTS, SCHWP, and WCC. The VTS in the observations with high ADs are smaller than the lower limits specified in each rule.

Further investigation shows that all these observations with high ADs are from Wednesday May 1, 2013, which is a public holiday in Hong Kong. The profiles of the NLTG, PAU and VTS power consumption on May 1, 2013 are shown in Figure 4.9.

The corresponding average power consumptions in Cluster 5 are also plotted for comparison. It is obvious that the measurements on May 1, 2013 are much lower than the corresponding average values. This case shows that the abnormality degree is effective in diagnosing non-typical and abnormal building operations.

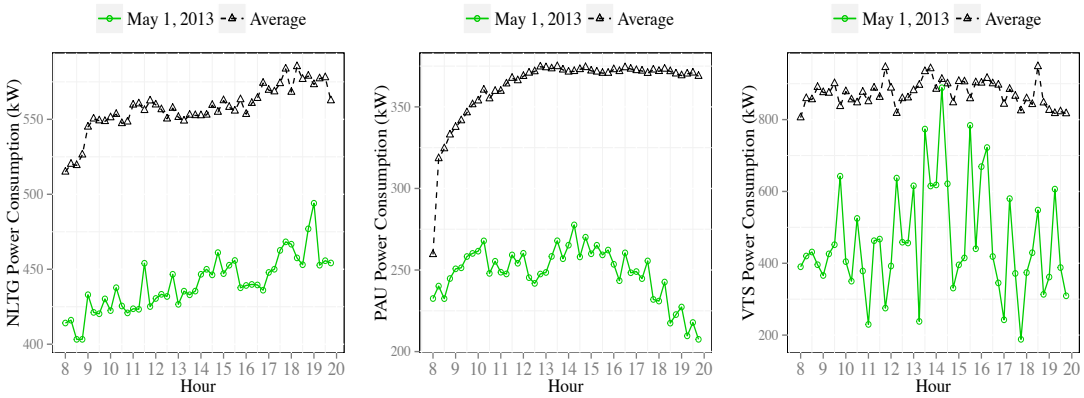


Figure 4.9 NLTG, PAU and VTS measurements on May 1, 2013 against the averages

Table 4.6 Summary of rules being violated

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

Furthermore, it is found that large ADs take place during the similar periods every day. Figure 4.10 shows the profiles of the means of ADs on Friday, Saturday, and Sunday. The profiles on weekdays are very similar, so only the profile on Friday is shown here as an example. There are two obvious spikes on Friday and other

weekdays as well, which take place during 6:00 to 9:00, and 19:00 to 21:00 Three main spikes are observed on Saturdays, and they are recorded during 7:00 to 9:00, 14:00 to 16:00, and 19:00 to 21:00. The profile on Sundays is relatively flat, and only one small spike is observed between 7:00 and 9:00. The results are in accordance with the domain expertise. The office hours for typical office buildings in Hong Kong are from 8:00 to 20:00 in weekdays, and 8:00 to 14:00 on Saturdays. The building system performs either a stage-up or a stage-down process during these periods. The transient changes normally result in very different operation behaviors. For instance, during the stage-up process, chilled water pumps usually consume much more power due to the motor starting characteristics. Figure 4.10 shows that a typical stage-up or stage-down process normally last for around 2 hours.

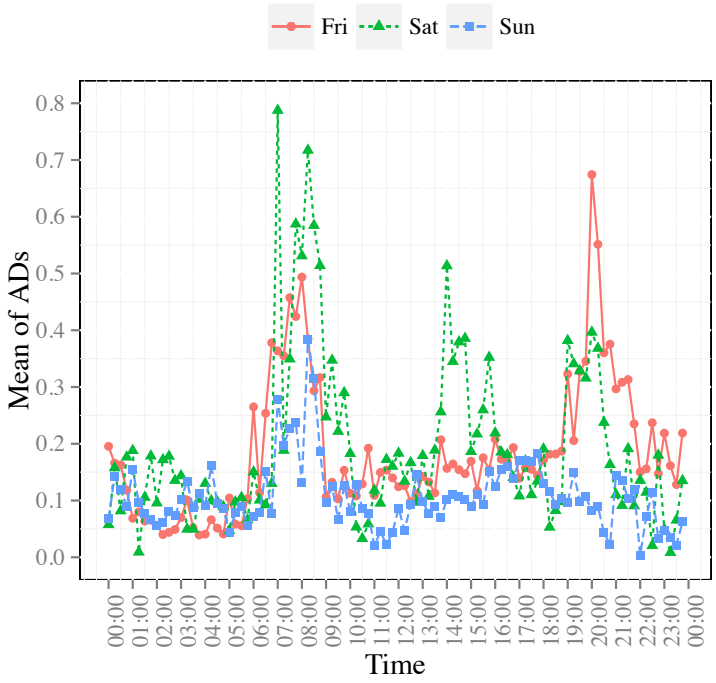


Figure 4.10 AD means on Fridays, Saturdays, and Sundays

4.4.3 Sensor Fault Diagnosis

It was found that the rules related to the VTS are frequently violated during the period between Mar 17, 2013 and Apr 22, 2013. Table 4.7 presents three examples of the rules being violated, which describe the associations between VTS and three main HVAC subsystems, i.e. WCC, PAU and SCHWP.

Table 4.7 Examples of the rules being violated related to VTS

No.	Antecedent (kW)	Consequent (kW)	Supp.	Conf.	Lift
1	WCC in [817.0, 1403.2]	VTS in [490.3, 1442.4]	0.29	0.96	1.48
2	PAU in [371.5, 406.9]	VTS in [565.5, 1614.5]	0.37	0.97	1.37
3	SCHWP in [60.4, 83.1]	VTS in [537.5, 1556.5]	0.31	0.96	1.36

The VTS power consumptions in the observations violating the rules are smaller than the lower limits of VTS in the rules. Figure 4.11 shows the VTS power consumption of the abnormal observations against those of the normal observations. It is shown the VTS power consumptions of abnormal observations are much lower. The power consumption of VTS in ICC consists of five parts, i.e., lifts in the car parking area, office shuttle lifts, office service lifts, fireman lifts, and escalators. Further investigation shows that during the above-mentioned period, the power meter for the office service lifts broke down and no value was recorded. Therefore, the aggregated power consumption for the VTS was smaller.

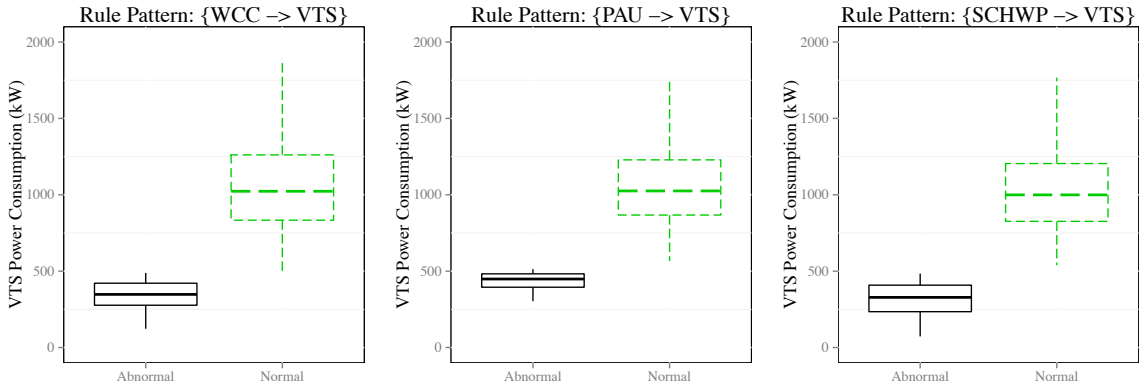


Figure 4.11 VTS power consumptions of normal and abnormal observations

4.5 Summary

This chapter described a methodology for mining cross-sectional knowledge in BAS data. The methodology is developed based on the generic DM-based framework proposed in Chapter 3. Considering the rich domain knowledge in the building field, unsupervised DM techniques are recommended as the primary means of discovering underlying data structures and relationships in BAS data. The methodology provides a reference for developing DM-based tools for cross-sectional knowledge discovery in massive BAS data and applications in building energy management.

The methodology has been implemented in analyzing the BAS data of International Commerce Centre, the tallest commercial building in Hong Kong. Both the association rule mining (ARM) and the quantitative association rule mining (QARM) are adopted to discover the cross-sectional associations in BAS data. BAS data are mainly numeric data. The implementation of ARM requires the data to be categorical, and therefore, a data discretization step becomes necessary. The discretization is typically performed based on domain knowledge. The results

obtained are usually more interpretable as the interval used for discretization is user-defined. However, improper data discretization can greatly degrade the quality and reliability of the knowledge discovered. The ARM-based methodology is validated using the ICC data. The knowledge discovered can be used to find abnormal behaviros in building operations and thereby, enhancing the building operational performance.

By contrast, the QARM method provides a more flexible way to discover associations from BAS data as it can directly work with numeric data. Nevertheless, the interval identified in the association rules are automatically generated and may not specifically meets the needs of building operation staff. Therefore, it may bring some difficulties in knowledge interpretation. Two indices of high practical values are defined to facilitate the post-mining of QARM, i.e. the standard deviation of lift (SD-Lift) of rules with similar rule pattern and the abnormality degree (AD). SD-Lift can help to fast select useful rules from a large number of rules obtained in ARM, which is a major obstacle to the application of ARM. AD provides a generic method of using the association rules for detecting abnormalities. These two indices are proven to be valuable for applying the knowledge discovered by DM (i.e. association rules in this case) to building diagnostics. The change of operation strategy, non-typical and abnormal operations and sensor fault occurring during operation in ICC are successfully detected and diagnosed. In practice, the selection between these two methods depends on the actual needs and the knowledge level of building

operation staff.

CHAPTER 5 DEVELOPMENT OF METHODOLOGY FOR TEMPORAL KNOWLEDGE MINING AND ITS APPLICATIONS

Considering that BAS data are in essence multivariate time series data, the cross-sectional knowledge discovered may not be able to fully capture the relationships over time. Building operations are typically dynamic due to the changes in indoor and outdoor operation conditions, such as the outdoor climate conditions, indoor occupant number and utilization of indoor electric appliances. Meanwhile, the changes hardly occur simultaneously which results that the dynamics in building operations are very complicated. For instance, the indoor temperature is influenced by the outdoor air temperature. However, when the infiltration is not significant, these two temperatures rarely change simultaneously due to building thermal mass. Time lags between them often bring challenges to the sequence control of chiller plants. The dynamics are usually complicated and have great influences on control performance, interactions among building components and integrations between buildings and communities (e.g., electricity power grid) [Xue et al. 2014]. In practice, it is desired to discover such temporal knowledge hidden in BAS data. Advanced tools and methods for temporal knowledge discovery should be developed for this purpose.

Conventional time series analytics, such as the autoregressive moving average models (ARMA), are mainly used for solving predictive tasks in the field of building management, including the prediction of building electricity consumption [Azadeh et al. 2008; Fernandez et al. 2011], building thermal load [Yao et al. 2004] and indoor environment [Yiu and Wang 2007; Zamora-Martinez et al. 2013]. In recent years, various approaches have been developed to mine temporal knowledge in different formats, such as events, clusters, motifs and temporal association rules [Madsen 2007; Fu 2011]. However, only limited studies have been performed to explore their potential in analyzing BAS data. Patnaik et al. adopted the motif discovery technique to mine chiller operation data in data centers [Patnaik et al. 2011]. Motifs (i.e., frequent sequential patterns) were successfully discovered to identify energy-efficient operation patterns. Miller, Nagy and Schlueter used a similar method to analyze building energy consumption data [Miller et al. 2015]. Energy consumption motifs were extracted for building performance characterization. Discords, or infrequent sequential patterns, were identified and used for fault detection. Their work demonstrated the encouraging potentials of time series data mining in the knowledge discovery of BAS data for managing building operations. Currently, the potential and applicability of various time series data mining techniques in mining big BAS data are still uncertain considering unique characteristics of BAS data, such as low quality, nonlinearity, multiple scales or units, and multicollinearity. A generic and systematic methodology for discovering temporal knowledge in big BAS data is needed for

developing applicable tools in BAS.

This chapter proposes a methodology for mining temporal knowledge hidden in big BAS data and demonstrates its applications in real cases. Section 5.1 presents the methodology developed for temporal knowledge discovery in BAS data. Section 5.2 describes the implementation of the methodology using a real-world BAS data. Section 5.3 illustrates the applications of the knowledge discovered in building energy management. The last section 5.4 summarizes this chapter.

5.1 Research Methodology

As proposed in Chapter 3, the DM-based analytic framework consists of four major phases, i.e., data preprocessing, data partitioning, knowledge discovery and post-mining. Each phase was specifically designed considering the BAS data quality and structure, data format requirement of DM techniques, interpretation and selection of knowledge discovered, and application of the knowledge to building performance assessment, diagnosis and optimization. The methodology presented in this chapter is developed within this framework, as shown in Figure 5.1. Three tasks are performed at the first phase, including data cleaning, period estimation and data transformation. Phase 2 adopts the evidence accumulation clustering to partition the SAX subsequences. Phase 3 adopts two techniques, i.e., motif discovery and temporal association rule mining, to discover two different types of knowledge. Two

post-mining methods are developed in Phase 4 to improve the efficiency and effectiveness of handling the large amount of knowledge discovered in Phase 3. The details of each phase are introduced in the following subsections.

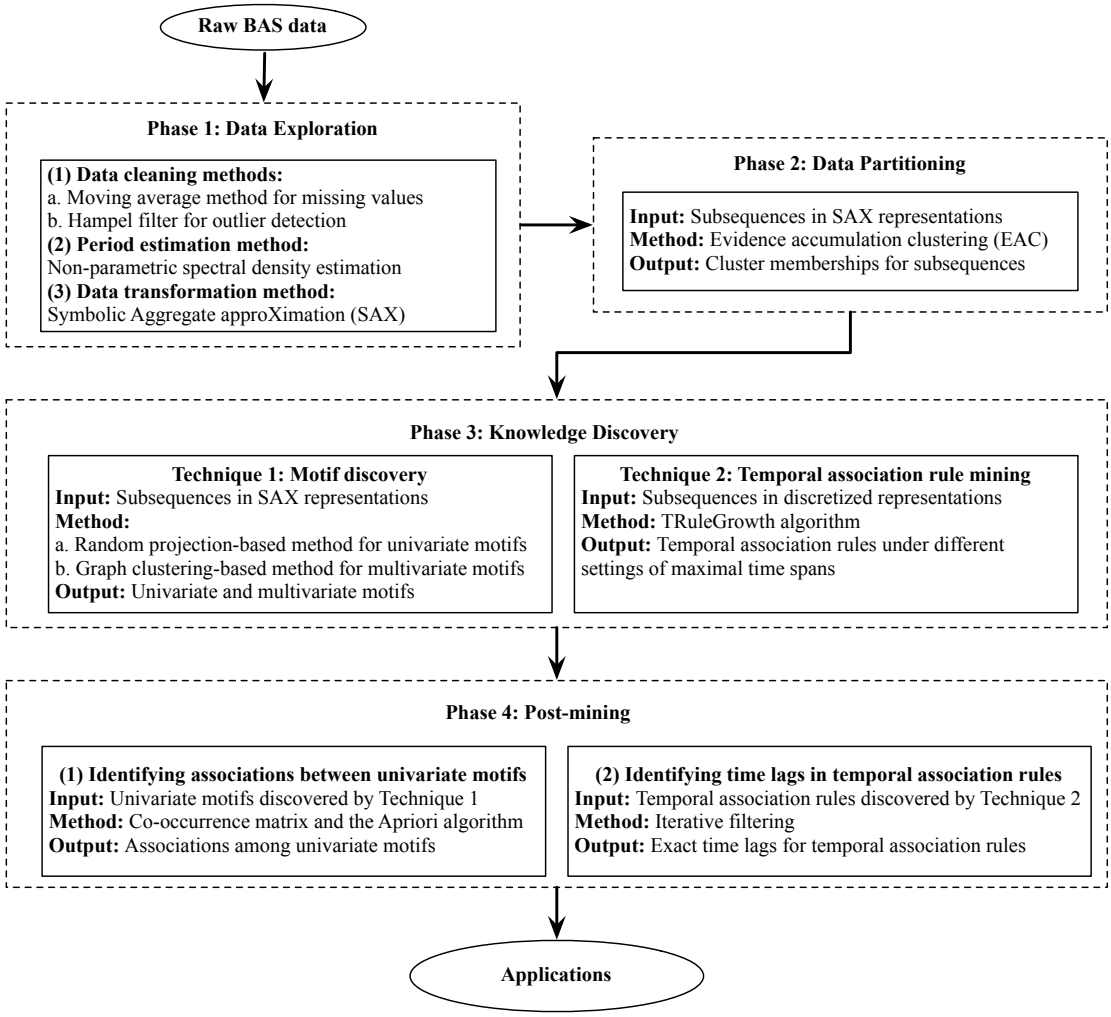


Figure 5.1 Research outline for temporal knowledge discovery in BAS data

5.1.1 Data Exploration

Data exploration fulfills three tasks, i.e., data cleaning, period estimation, and data transformation, with the aims to enhance the data quality, explore the intrinsic characteristics in BAS time series data, and prepare the raw data with suitable format

for data mining.

5.1.1.1 Data Cleaning

Data cleaning aims to improve BAS data quality by filling missing values and detecting outliers in raw BAS time series data. Outliers in a time series are observations that are highly unlikely to occur based on the variation seen in the rest of the time series. They can be classified into two types, i.e., points as outliers and subsequence as outliers [Gupta et al. 2014]. This phase only handles the first type of outliers in the raw time series data, as the identification of the second type of outliers may overlap with mining discords (i.e., infrequent sequential patterns) in the later process. Similar to the methodology proposed in Chapter 4, the Hample filter is applied for outlier detection. The moving window-based method is used to impute the missing values with a short duration, i.e., less than 2-hour. Any missing values with a longer time duration are excluded from analysis.

5.1.1.2 Period Estimation

This step is specifically developed for time series data, considering that long time series data usually exhibit periodicity, and consequently motifs and association rules periodically repeat. Finding the period in the time series data and then segment those data into short subsequences can considerably reduce the mining load. It is a

common practice in time series data mining, particularly in handling very long time series data like BAS data. The repeating daily working schedule of building users (e.g. office hours and non-office hours) results that the operation schedules of major systems and equipment (such as air conditioning, lighting and lift systems) usually repeat daily. Obviously, the BAS data exhibit daily periodicity. In view of this, Miller, et al. segmented the time series of building energy consumption data into daily sequences in their study [Miller et al. 2015]. This research attempts to adopt a data-driven approach to estimating the intrinsic periods embedded in BAS data. There are two purposes for doing this: firstly to minimize the dependence on domain knowledge in the knowledge discovery process; secondly to maximize the possibility of discovering new knowledge, or new periods in BAS data in our case. Periods in time series data can be detected using the spectral density estimation methods, which can be either parametric or non-parametric. The parametric methods first model the time series using time series modeling techniques, such as autoregressive and moving average (ARMA). The spectral density is then estimated based on the model parameters. By contrast, the non-parametric methods estimate the spectral density by taking the Fourier transformation of the autocorrelation function. Considering that the building data usually present diurnal, weekly and annual seasonality, the resulting parametric models can be very complex. Therefore, this study applies the non-parametric method to period estimation.

5.1.1.3 Data Transformation

Data transformation prepares the time series data with suitable formats to meet the following two needs. Firstly, different mining techniques require different data formats (e.g., numerical or categorical) and BAS data exhibit diversity in units, scales, and data types. Secondly, the computation load is a big concern due to the huge volume of big data, which can be alleviated by effectively reducing the volume of the data without losing valuable information embedded in the data. In this study, the symbolic approximation aggregate (SAX) method is proposed to transform the original time series BAS data into meaningful symbols [Lin et al. 2007; Miller et al. 2015]. The SAX method transforms a numeric time series into a symbol stream and the length of the symbol stream is much shorter than the original time series. It can therefore reduce the data size.

To perform SAX, a univariate time series of length n is firstly standardized to have a zero mean and a standard deviation of 1 and then segmented into m subsequences with a window size of q. One of the typical methods to segment the time series is based on the period detected in the previous step. For example, if the period estimated is 24 hours, one day BAS data will form one subsequence. Two parameters need to be defined to perform SAX, i.e., the word size W and the alphabet size A. A set of breakpoints (e.g., $\beta_1, \beta_2, \ldots, \beta_{A-1}$) are determined in such a manner that the area under the *N(0,1)* Gaussian curve from β_i to β_{i+1} is $\frac{1}{A}$. Each interval will be assigned with an alphabet (e.g., a, b, and c) and the number of alphabets used

is the alphabet size, A. Given the word size (W), each subsequence in the window size of q can be divided into W equal sections, and the means of each sections are calculated. According to which interval (i.e., β_i to β_{i+1}) the mean lies within, the corresponding alphabet is assigned to the section. In this way, each subsequence can be represented by a SAX word which consists of W alphabets. For example, *abca*, *aabc*, *bcca* are SAX words given $W=4$ and $A=3$. In these SAX words, the alphabet size (A) is 3, so three alphabets (i.e., a, b and c) are used; the word size (W) is 4, so each SAX word consists of four alphabets. The original time series is transformed into a string of alphabets. The larger the alphabet size (A) and the word size (W), the more detailed information retained in the symbolic stream. However, the reduction of computation load becomes less. Therefore, there is a trade-off, which will be discussed in the later case studies.

The distance between two SAX representations are calculated as $\sqrt{\frac{q}{w}} \times \sqrt{\sum_{i=1}^{w} dist(S_i, B_i)^2}$, where S and B are two SAX representations, and *dist()* is the distance function for SAX symbols. Table 5.1 presents an example of distance matrix between symbols considering an alphabet size of 4. The value in *cell(x,y)* is calculated using Equation 5.1. A dissimilarity matrix considering different SAX representations can be computed accordingly. More details can be found in [Lin et al. 2007].

Besides SAX, difference-based and dictionary-based methods are also capable of transforming time series into symbols [Daw et al. 2003; Kwac et al. 2014; Gulbinas et al. 2015]. The difference-based method transformed the raw time series into symbols

based on their first- or higher-order differences. It can be used when the changes between successive time steps are more important than the absolute values [Daw 2003]. The dictionary-based methods transform the time series into symbols by matching the raw data with predefined patterns in a dictionary. For instance, in the studies performed by Kwac et al. [Kwac et al. 2014] and Gulbinas et al. [Gulbinas et al. 2015], clustering analysis was applied to generate the representative patterns of daily power consumption, based on which a dictionary was built for symbolization. SAX is selected in this study considering the following two aspects. Firstly, SAX is straightforward to use, as it requires little domain expertise and preprocessing. Secondly, SAX contains an intrinsic distance measure, which provides extra value in the subsequent knowledge discovery [Lin et al. 2007], as shown in the later part.

$$cell(x,y) = \begin{cases} 0 & if\ |x-y| \le 1 \\ \beta_{\max(x,y)-1} - \beta_{\min(x,y)} & otherwise \end{cases} \quad (5.1)$$

Table 5.1 An example distance matrix for SAX symbols

Distance	*a*	*b*	*c*	*d*
a	0	0	0.67	1.34
b	0	0	0	0.67
c	0.67	0	0	0
d	1.34	0.67	0	0

5.1.2 Data Partitioning

Due to the changing operation conditions and complicated system dynamics and

interactions, the big BAS data usually scatter in a high-dimensional space. To enhance the reliability and sensitivity of the mining results, data partitioning is carried out to divide the data into several groups or clusters, with the aim of maximizing the intra-group similarities while minimizing the inter-group similarities. Knowledge discovery are then performed on each group separately. Clustering analysis is a suitable DM technique to perform this task. Despite of the large number of clustering algorithms being available, no single algorithm is able to identify all kinds of cluster shapes and data structures in practice [Fred and Jain, 2005]. It is usually very difficult to find out the optimal clustering algorithm and the settings of its parameters. Some methods have been developed to facilitate the decision-makings, based on either internal (e.g., Dunn index and Davies-Bouldin index) or external validation indices (e.g., purity and mutual information). However, no validation method can impartially evaluate the results of any clustering algorithm [Vega-Pons and Ruiz-Schulcoper 2011]. A common practice is to try out a large number of algorithms with different parameters in order to obtain desired the clustering results. The process can be computationally expensive and time-consuming.

Ensemble learning is capable of enhancing the clustering performance by combining a number of base learners, whose individual performance may be poor [Fred and Jain, 2005; Vega-Pons and Ruiz-Schulcoper 2011]. The evidence accumulation clustering (EAC) is a method designed to apply ensemble learning on clustering analysis [Gupta et al. 2014]. One advantage of the EAC over other

conventional clustering methods is that it has the ability to discover clusters with various sizes and shapes. In addition, the method can automatically determine the optimal cluster number, which provides great flexibility in analyzing data with unknown characteristics. The partition around medoids (PAM) is selected as the base algorithm for EAC. PAM shares a similar partitioning mechanism as the popular *k*-means algorithm. Compared to the *k*-means, PAM is more robust to outliers and noises and can take a dissimilarity matrix as inputs. Therefore, PAM is more compatible with time series data in SAX representations.

Three parameters needs to be defined to perform EAC, i.e., the total iteration number E, the lower and upper limits of the cluster number K_{lower} and K_{upper}. E sets of clustering results are generated by PAM with different cluster numbers (i.e., randomly sampled from K_{lower} to K_{upper} in each iteration) and the dimension of input data. These E sets of clustering results are then transformed into a co-occurrence matrix. Assuming that the data contains n observations, the co-occurrence matrix C has a dimension of $n \times n$. The value of $C_{i,j}$ is the number of times when observations i and j are grouped in the same cluster divided by the total iteration number E. The final clustering result is obtained by using hierarchical agglomerative method to cluster the co-occurrence matrix.

5.1.3 Temporal Knowledge Discovery

After the data are preprocessed and partitioned, appropriate DM techniques will be applied for knowledge discovery. The typical descriptive knowledge types in time series data include motifs, discords and temporal association rules [Fu 2011].

Motif Discovery

Motif, or frequent sequential pattern, is a typical knowledge type which can be discovered in time series data. Motifs are valuable to temporal association rule mining, discord (i.e. infrequent sequential pattern) detection, and time series classification [Chiu et al. 2003].

Motif discovery has been mainly applied to analyze univariate time series in previous studies. Conventional motif discovery methods are based on exhaustive search, which results that the computational costs increase dramatically for long time series and is therefore not applicable to big data. In view of this, a more efficient algorithm, which is based on random projection and compatible with SAX representations [Chiu et al. 2003], is selected to discover univariate motifs. Assuming that the time series has a length of n and the sliding window size is q, a matrix containing all the subsequences (denoted as M_1) can be constructed and has a dimension of $(n-q+1)\times q$. Each subsequence is transformed into a SAX representation. Assuming the word size is W, the new matrix containing the SAX

representations (denoted as M_2) has a dimension of $(n-q+1)\times W$. Random projection is performed by randomly picking *s* columns from M_2, where *s* ranges from 1 to *W*-1. A collision matrix, which has a dimension of $(n-q+1)\times(n-q+1)$, is constructed to record the times of being identical for two subsequences after a number of random projections. A tentative univariate motif is identified if the two subsequences result in a high value in the collision matrix. Potential members of this tentative univariate motif can then be identified by calculating the Euclidean distance in the original numeric representations.

Several methods have been developed to identify motifs in multivariate time series data, such as PCA-based and density estimation-based methods [Tanaka et al. 2005; Minnen et al. 2007]. Those methods can successfully identify synchronous multivariate motifs. However, their practical value in analyzing real-world data is limited, as the motifs in multivariate time series data do not necessarily start at the same time and their duration may vary as well. We can see a lot of such examples in building operations. For example, when the air conditioner or chiller is turned on, the indoor temperature will not change immediately due to the thermal mass. The sudden increase of the lift power consumption in the morning peak hour does not correspond to a large increase in the chiller power consumption due to the pre-cooling strategy. In this research, multivariate motif discovery algorithm proposed in [Vahdatpour et al. 2009] is adopted. The main advantage is that, firstly, both synchronous and non-synchronous multivariate motifs can be discovered, and secondly, the

multivariate motifs identified may consist of all univariate motifs or any subset of the univariate motifs. The method first performs univariate motif discovery on the time series of each variable. A graph clustering approach is then applied to identify multivariate motifs. A directed coincidence graph G is constructed. Each motif r_i is represented by a node v_i. $e_{i,j}$ represents the edge connecting the node v_i and v_j. The weight of $e_{i,j}$ is denoted as $w_{i,j}$ and calculated as $coincident(r_i, r_j)/size_i$, where $coincident(r_i, r_j)$ is the total number of times that a temporal overlap is found between r_i and r_j and the $size_i$ is the number of occurrence of r_i. A parameter, α, ranging from 0 to 1, is user-specified as the minimum correlation between univariate motifs based on which a multivariate motif could be constructed.

Temporal Association Rule Mining

The difference between association rule mining (ARM) and temporal association rule mining (TARM) lies in whether the temporal information is contained in the rule or not. ARM was mainly used to discover cross-sectional associations, where the temporal information is neglected. The typical format of ARM is $A \rightarrow B$, where $A \cap B = \emptyset$. It states that if A happens, B will also happen. An association rule is derived if both the rule support and confidence exceed the user-defined thresholds. The support of a rule is the fraction between the number of times when both the antecedent and consequent take place and the total number of records. The confidence of a rule is the conditional probability of the consequent given the antecedent. The interestingness of the association rules can be evaluated using the *lift*, which is the

ratio between the rule confidence and the support of consequent. It measures the dependency and correlation between the antecedent and the consequent of a rule. Potentially useful rules usually have a *lift* larger than 1, indicating that the occurrence of the antecedent positively influences the occurrence of consequent.

Temporal association rule mining (TARM) is of particular interest in mining BAS data because of the complicated dynamics in building operations. TARM, or sequential rule mining, discovers associations among variables while providing an insight into the temporal dependency. The general format of temporal association rules is also $A \rightarrow B$, where $A \cap B = \emptyset$. However, the temporal dependency is contained, indicating that B will take place after A. Various algorithms have been developed for deriving temporal association rules, such as the SPADE and CMRules [Zaki 2001; Fournier-Viger et al. 2012]. In engineering practice, temporal rules that are valid within a limited time span are of special interest. The format of such temporal rules is $A \overset{t}{\rightarrow} B$, which means that B will occur within t time units after the occurrence of A. Therefore, the TRuleGrowth algorithm, which can derive temporal association rules under the constraint of maximum time span [Fournier-Viger et al. 2012], is selected in this study. To perform this algorithm, three parameters need to be defined, i.e., the minimum support, minimum confidence, and the maximum time span. The other advantage of the TRuleGrowth algorithm is that it can greatly reduce the number of rules generated by controlling the maximum time span. Consequently, the post-mining phase consumes much less time.

5.1.4 Post-mining

The post-mining phase aims to build a bridge between knowledge discovered at Phase 3 and practical applications, such as building performance assessment, fault diagnosis and optimization. It usually needs domain expertise to select, interpret and apply the knowledge discovered. The process can be very time-consuming, due to the large amount of knowledge discovered and the diversity of knowledge representations (e.g., rules, clusters, decision trees). Application of the motifs and temporal association rules is straightforward. They can be used as the references for normal operations and anomalies can be detected if building operation patterns are different from those frequent patterns or violate the association rules. In this study, two methods are specifically developed to enhance the efficiency in post-mining and maximize the practical values of temporal knowledge discovered.

Identify Associations between Univariate Motifs

Building operations involves multiple separate and interactive subsystems, such as air conditioning, mechanical ventilation, lift, lighting and security systems. Univariate motifs usually represent the frequent sequential operation patterns of each system. It is reasonable to link the associations among univariate motifs with the interactions among subsystems. Multivariate motifs can provide general information

on which univariate motifs frequently occur together. However, they hardly quantify the relationships among univariate motifs and this limits their practical value. For example, a multivariate motif cannot answer, if one univariate motif occur, whether the other univariate motifs in it will occur or not with certain probability. In this study, a post-mining method is designed to explore the associations among univariate motifs which can directly answer this question. This method is an extension of association rule mining. Given a multivariate time series data, the univariate motif discovery algorithm is applied to each univariate time series separately to find univariate motifs. These univariate motifs are then labeled as $m_1, m_2, \ldots, m_L$, where L is the total number of univariate motifs discovered. Afterward, a co-occurrence matrix is constructed. The matrix has L columns. The values of each row are either 1 or 0, indicating whether an occurrence of a univariate motif is observed or not. Once the matrix is constructed, the Apriori algorithm is used to discover associations between univariate motifs. Two parameters, i.e., the minimum thresholds for support and confidence, are defined for rule induction. Three statistics, including the support, confidence and lift, can be generated with each association rule to facilitate decision making.

An example for construction of a co-occurrence matrix is given here. Figure 5.2 illustrates five univariate motifs (i.e., m_1 to m_5) in the sequences of three variables, A, B and C. The motifs in A and B, m_1 to m_4, have a time duration of 10 while m_5 in C has a time duration of 8. The co-occurrence matrix is constructed as shown in Table 5.2. The numbers (0 or 1) in each row show the occurrence of the corresponding

motifs. For example, the second row shows that only motif 5 occurs during the time period between 18 to 25; the first and third rows show that motifs 1 and 3 occur together twice; the fifth row shows that motifs 2, 4, 5 occur together for once; the sixth row shows that motifs 2, 3 and 5 occur together for once. It should be noted that, although the occurrence of the motifs are related to certain time period, the exact time is not considered in constructing the matrix. The frequency of the co-occurrence of multiple univariate motifs is of interest.

The construction of the co-occurrence matrix can be conveniently implemented by programming with the information of starting and ending time instants of all univariate motifs. Once the co-occurrence matrix is ready, the Apriori algorithm is adopted to mine the associations. Setting the minimum thresholds of support and confidence as 0.3 and 0.8 respectively, two rules are derived, i.e., $m_1 \rightarrow m_3$ and $m_2 \rightarrow m_5$. Both rules have a support of 0.4 and a confidence of 1. It means that when motif 1 occurs, the probability of the occurrence of motif 3 is very high.

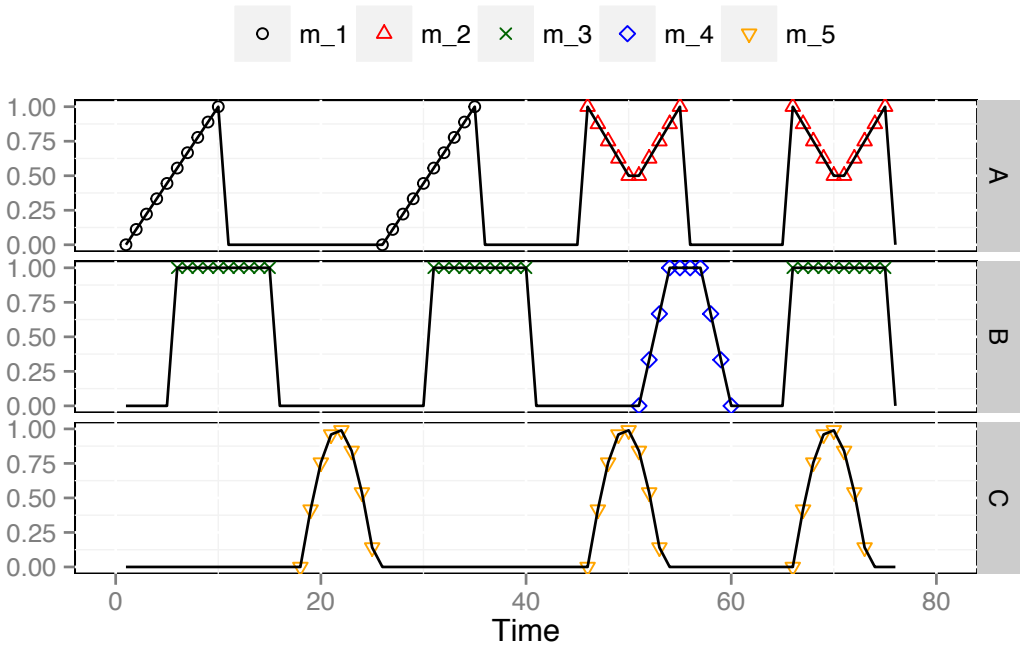


Figure 5.2 An example of univariate motifs discovered in three dimensions

Table 5.2 An example of co-occurrence matrix for mining association rules between univariate motifs

m_1	m_2	m_3	m_4	m_5
1	0	1	0	0
0	0	0	0	1
1	0	1	0	0
0	1	0	1	1
0	1	1	0	1

Identify Time Lags in Temporal Association Rules

The TRuleGrowth algorithm is adopted to discover the temporal association rules under the constraint of a maximum time span. One limitation is that no information is available about the exact time lag, which is the time interval between

the antecedent and the consequent. This type of information is valuable for establishing reliable control and performance optimization in building operations. An iterative filtering method is developed to identify the time lag. The method iteratively runs the TRuleGrowth algorithm by changing the maximum time span from 1 to T and the temporal association rules generated at each iteration and the corresponding time lag are stored in the rule sets. The time lag in a temporal association rule can be discovered by matching the rule with the rule sets.

5.2 Mining Real BAS Data

5.2.1 Identification of Daily Power Consumption Patterns in Building Operations

The methodology is applied to the BAS data retrieved from ICC. The intrinsic periods in the time series of building total power consumption are estimated using the non-parametric spectral density estimation method. The top three dominant frequencies are 0.0103, 0.0417 and 0.1121, which correspond to periods of 97 (i.e., 1/0.0103), 24 (i.e., 1/0.0417) and 9 (i.e., 1/0.1121) respectively. Since the BAS data are collected at an interval of 15-minute, these three periods are approximately 1-day, 6-hour and 2-hour respectively.

The dominant period in the sequence of building total power consumption is 1-day. Therefore, the whole BAS data are segmented into daily subsequences and

then transformed into SAX representations. Increasing the word size W and alphabet size A will lead to a better SAX representation of the original time series. However, the reduction in computation load is less. Miller et al. recommended W and A as 4 and 3 respectively to identify typical patterns in building power consumption data [Miller et al. 2015]. Actually, the selection of W and A is influenced by the scale of the building, installation capacities (e.g., cooling, heating, total electricity power) and operation strategies. A large building with high installation capacities tends to require large W and A to adequately describe the variation in the original time series data. In this study, W is chosen as 12, considering that 2-hour was identified as one of the dominant periods. Considering that the chiller plant usually accounts for a large proportion of the total power consumption and the maximum running chiller number is 5 in the BAS data to be analyzed, A is chosen as 5 to reflect there are five major levels of power consumption due to the on-off control of chillers. It should be noted that the standardization is only applied to the total building power consumption time series, but not the daily subsequences. The consideration here is to identify typical daily patterns considering both the shape and magnitude.

The SAX representations of daily subsequences are then partitioned into different groups using the EAC method. K_{lower} and K_{upper} are selected as 2 and 20 respectively. The iteration number E is set as 200. As a result, 8 clusters are identified. Clusters 5, 6, 7 and 8 only consists of 6 daily subsequences out 365 subsequences. Those subsequences are actually subsequence-wise outliers, as their shape and

magnitude are dramatically different from the others. They are excluded from further analysis. Figure 5.3 presents the profiles of daily subsequences in Clusters 1 to 4. Further examination of each cluster shows that Clusters 1 to 4 can be best interpreted using the climate and day type. Cluster 1 includes weekends in cold season and Cluster 4 contains weekdays in hot season. Cluster 2 and Cluster 3 mainly include weekdays in cold season and weekends in hot season respectively. The clustering results are coincident with the results obtained in Chapter 4 and domain knowledge. It indicates that the SAX transformation can very well preserve the important information in original time series data.

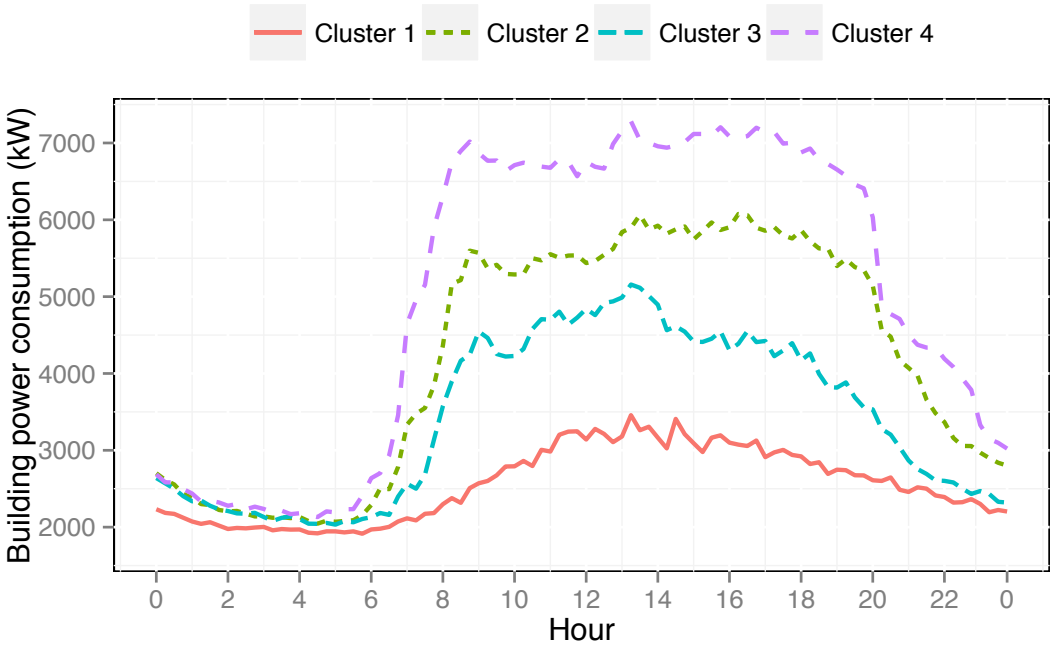


Figure 5.3 Four typical prototypes of daily building power consumption

5.2.2 Identify Frequent Operation Patterns of Subsystems

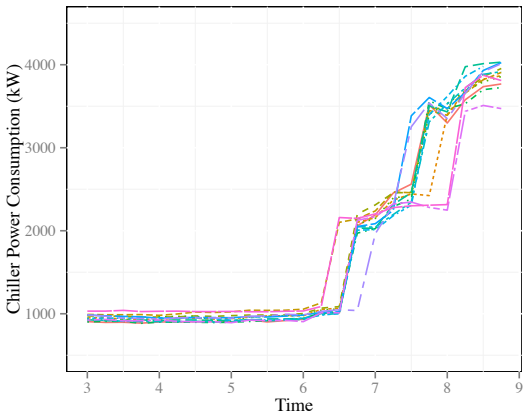
Univariate and multivariate motif discovery are applied to the 4 clusters separately to identify the frequent operation patterns. Considering that the daily operation conditions (including outdoor weather conditions and indoor occupancy and equipment utilization conditions) varies largely, it is more meaningful to discover motifs in building operations with smaller lengths, compared with the above identification of power consumption pattern. In this study, the length of the univariate motifs to be discovered is set as 6-hour, as it is identified as the second dominant period in the building power consumption data. More specifically, subsequences are segmented using a 6-hour sliding window, which means the subsequences created are overlapping. Standardization is performed for each subsequence in each cluster. SAX representations are created using the setting of W=6 and A=5. In such a case, each SAX symbol represents the hourly mean and has five possible levels. The iteration number for random projection is 100. During each iteration, 4 out of 6 SAX symbols are randomly selected for comparison, which means that subsequences belonging to the same motif can be different at one position at most [Chiu et al. 2003].

Table 5.3 summarizes the number of univariate motifs discovered for each subsystem in Cluster 4 (i.e., weekdays in hot season). Figure 5.4 presents 4 motifs discovered in the time series of the aggregated chiller power consumption in Cluster 4. Each curve represents an occurrence of the corresponding motif. It is apparent that the time series subsequences belonging to the same motif are very similar in their shapes

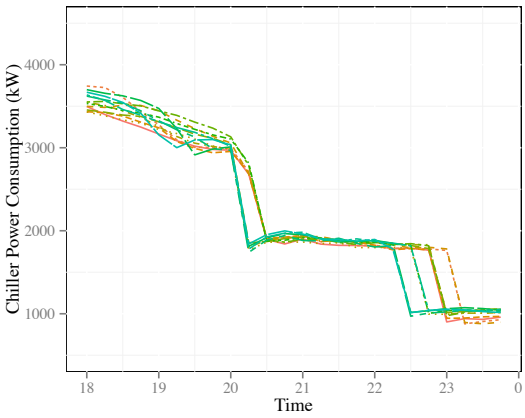
and magnitudes. An uptrend in chiller power consumption is observed in Figure 5.4a. It is shown that two chillers are sequentially switched on at the beginning of working hours (i.e., 6:00 to 9:00) to cope with the upcoming morning peak of occupancy and equipment utilization. The chiller switch-off process shares a similar pattern and two chillers are sequentially switched off (Figure 5.4b). The other two motifs, as shown in Figures 5.4d and 5.4c, present relatively steady operation conditions. The chiller operation between 0:00 and 6:00 is steadily maintained at a low level due to the absence of occupancy. By contrast, the chiller power consumption is maintained at a much higher level between 9:00 to 15:00. A slight decrease can be observed from 13:00 to 14:00, which is in accordance with the lunch time for most companies in ICC.

Typical operation behaviors can be obtained by analyzing the univariate motifs identified. For instance, Figure 5.5 presents 2 frequent patterns for the AHU operation between 21:00 and 3:00 in Cluster 4. The main difference is that a sudden drop in AHU power consumption is observed at 12:00 in Figure 5.5a, while the AHU power consumption gradually decreases in Figure 5.5b. After carefully examined the original data, it is found that the AHU power consumption measured at three mechanical floors (i.e., 6/F, 42/F and 78/F) simultaneously drop at 12:00 in pattern 1. By contrast, the drops are observed at 22:00, 1:00 and 2:00 for the AHUs at 42/F, 78/F and 6/F slightly and gradually in pattern 2.

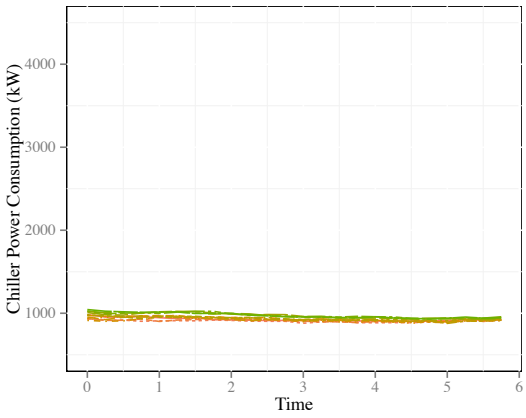
Table 5.3 A summary of univariate motifs discovered in Cluster 4



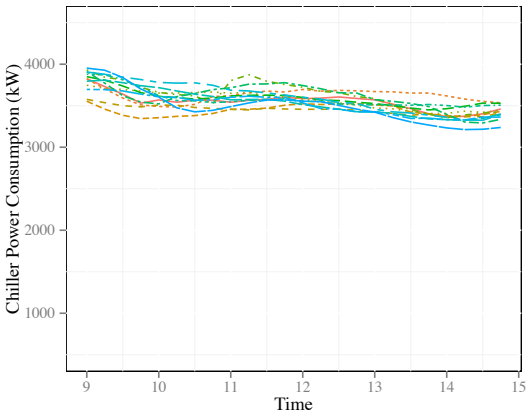
(a) Uptrend in chiller operation



(b) Downtrend in chiller operation



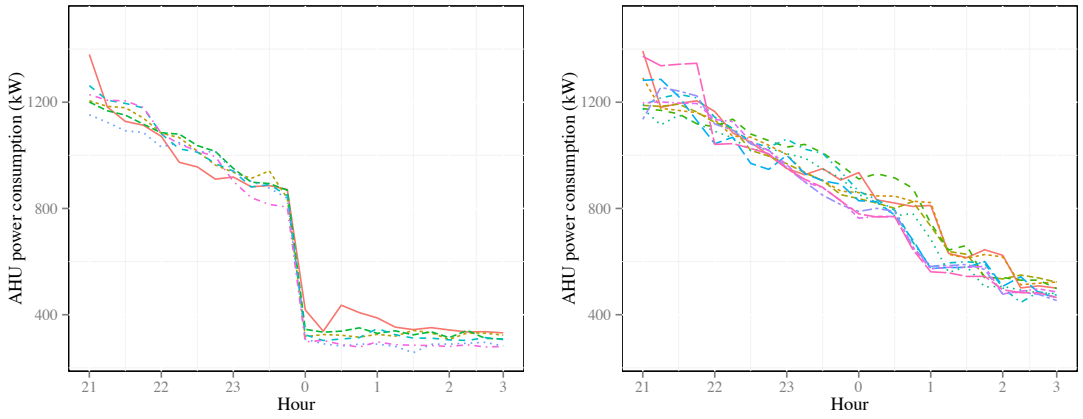
(c) Horizontal trend at low level



(d) Horizontal trend at high level

Figure 5.4 Examples of univariate motifs in chiller operation in Cluster 4

Subsystems	Chiller	CT	SCHWP	AHU	PAU	MV	VTS	NP	EP	PD
Motif No.	15	9	10	17	14	19	4	15	3	3

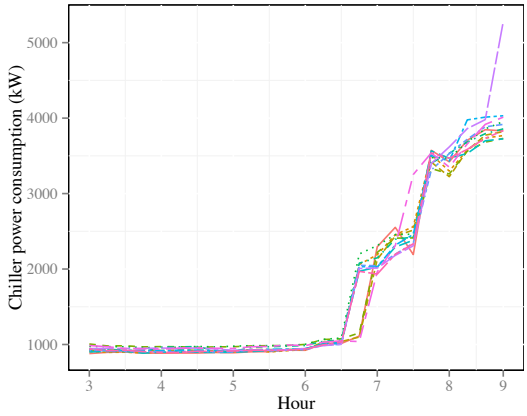


(a) Pattern 1 (b) Pattern 2

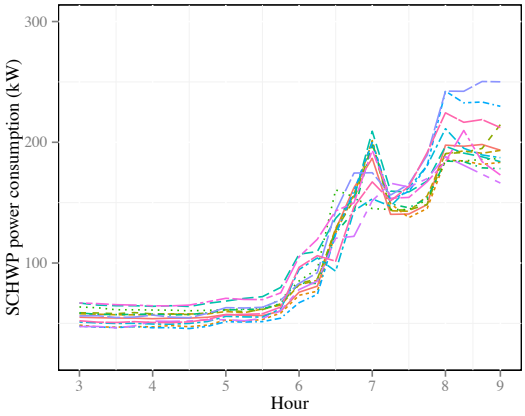
Figure 5.5 Typical AHU operation between 21:00 and 3:00 in Cluster 4

As introduced in Section 5.1.3, univariate motifs discovered are used to identify multivariate motifs. The algorithm can discover both synchronous and non-synchronous multivariate motifs. The parameter α is set as 0.8. Figure 5.6 presents an example of simultaneous multivariate motif. It depicts the building dynamic operations for different subsystems during 3:00 to 9:00. The chiller power consumption starts to rise from 6:30 and two chillers are sequentially switched on. A rise in SCHWP power consumption can be observed accordingly to circulate the chilled water. The PAU power consumption stays steady at low-level until 8:00. This is because ICC adopts demand-controlled ventilation to control the PAU and the occupancy increases from 8:00 because people start to work. A rise in MV power consumption can be observed at 8:00, which is also due to the occupancy change. In addition, the MV power consumption also undergoes an increase at around 6:30, which relates to the activation of the precooling strategy. Similarly, uptrends in VTS

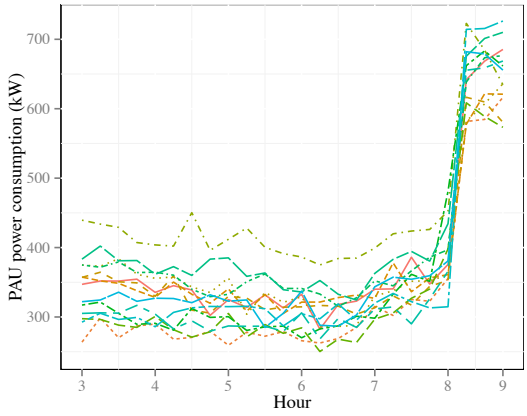
and NLTG power consumptions can be observed in Figures 5.6e and 5.6f to cope with the increase in occupancy. These motifs show that the HVAC system in ICC is under reliable control and operations well meet the expectations. ICC was awarded as an Intelligent Building of 2011 by the Asian Institute of Intelligent Buildings, partly owing to the advanced BAS installed in ICC.



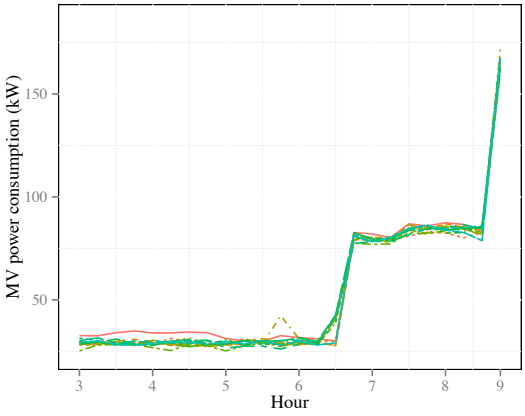
(a) Motif in chiller power consumption



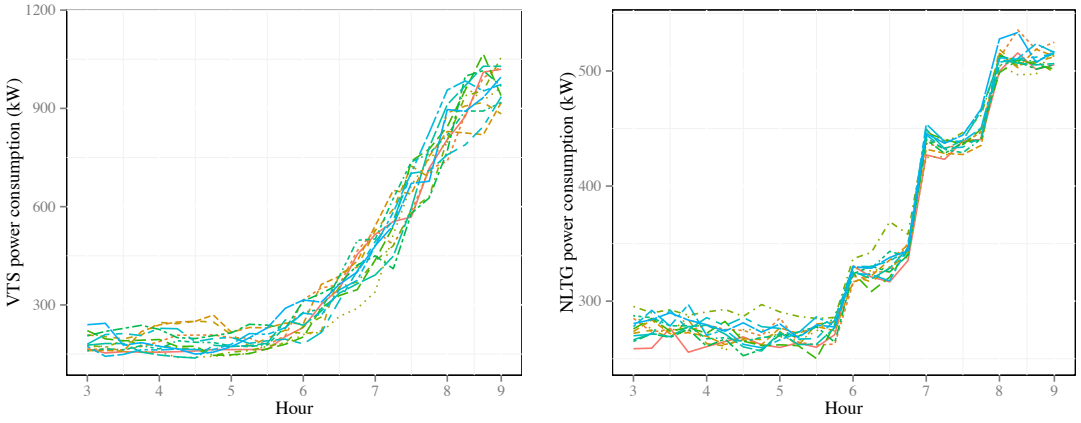
(b) Motif in SCHWP power consumption



(c) Motif in PAU power consumption



(d) Motif in MV power consumption



(e) Motif in VTS power consumption (f) Motif in NLTG power consumption

Figure 5.6 An example of multivariate motif in Cluster 4

5.2.3 Identify Temporal Associations between Subsystem Operations

This section focuses on discovering the temporal associations between the operations of different subsystems. The operation of each subsystem at certain time instant is represented by two features, i.e., level and trend. The power consumption data of each subsystem are categorized into three levels, i.e., *Low*, *Medium* and *High*. The trend is defined based on the changes between successive time step and categorized into 1 to 7, indicating large decrease, moderate decrease, slight decrease, steady, slight increase, moderate increase and large increase. The categorization thresholds are determined using k-means clustering algorithm. The TRuleGrowth algorithm is applied with the minimum support and confidence being set as 0.2 and 0.8 respectively. The maximum time span changes from 1 (i.e., 15-minute) to 12 (i.e., 3-hour). The post-mining method described in Section 2.4.2 is applied to find the

exact time lag in temporal association rules.

Table 5.4 presents three example rules describing the inter-subsystem temporal associations in the multivariate motif shown in Figure 5.6. The first rule shows that when the AHU power consumption is *Low* and experiencing a slight increase at time *T*, the chiller power consumption will be *Low* and stay steady at time *T+1*. The second rule shows that given the same antecedent, a slight increase in the chiller power consumption will be observed at *T+2*. These two rules demonstrate that the change in AHU and chiller operation is not synchronous and the time lag is around 15 minutes. The last example rule describes the temporal association between the NLTG and the PAU power consumptions. It states that when the NLTG consumption is *Low* and experiencing a significant increase at time *T*, a significant increase in the PAU power consumption will be observed at *T+9*. The result's validity can be verified by manually inspecting Figure 5.6. For instance, the first significant increase in NLTG and PAU power consumptions take place at around 5:45 and 8:00 respectively and therefore, the time lag for the third rule should be 9 unites of time (i.e., 135 minutes).

Table 5.4 Examples of temporal associations discovered

Rule	Antecedent	Consequent	Time lag (15-min)	Supp.	Conf.
1	AHU=Low, 5	Chiller=Low, 4	1	1.00	1.00
2	AHU=Low, 5	Chiller=Low, 5	2	0.89	0.89
3	NLTG=Low, 7	PAU=Low, 7	9	0.78	0.82

5.3 Applications of Temporal Knowledge Discovered

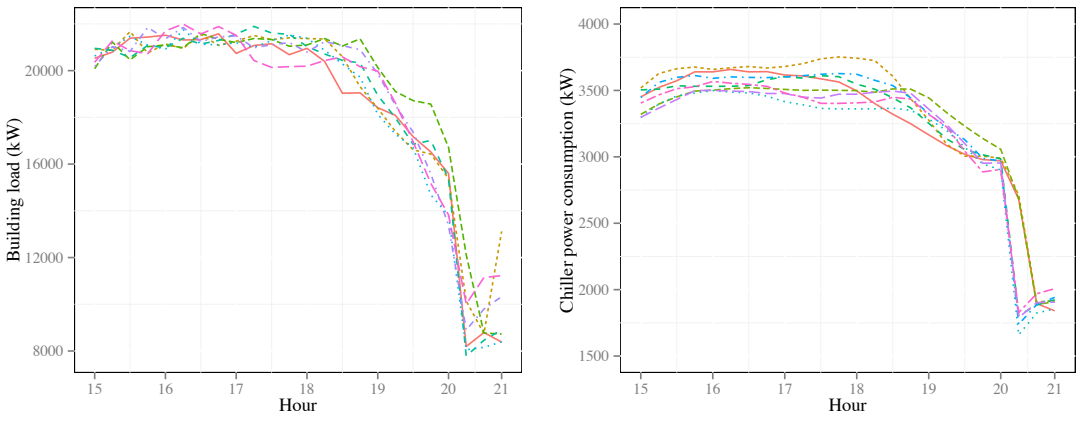
A straightforward approach to applying the temporal knowledge discovered to building management is to build a database of motifs and temporal association rules as the benchmark of building operations. Then, the real-time BAS time series data are compared with the benchmarked operations to identify any possible anomalies. The post-mining methods developed in this study provide two more approaches to such applications. The following parts demonstrate these applications.

5.3.1 Applications of Associations between Univariate Motifs

The post-mining method introduced in Section 5.1.4 is applied to discover associations between univariate motifs. To illustrate, 103 univariate motifs which are discovered in Cluster 4 are used for analysis. The Apriori algorithm is applied with the minimum support and confidence set as 0.1 and 0.8 respectively. These thresholds are set in such a way to ensure the discovery of strong but not necessarily frequent associations. 144 association rules are discovered. The association rules obtained can be applied to find anomalies in operation, such as less energy-efficient operations, faulty operations, as well as normal but rare operations.

As shown in Figure 5.7, one rule is *Cooling Load = Motif 3* $\rightarrow$ *Chiller = Motif 11*. It describes the association between Motif 3 in the building cooling load and Motif 11 in the chiller power consumption, which both take place between 15:00 to

21:00. Atypical patterns are identified by finding the time series data which meet the antecedent but not the consequent. An example is presented in Figure 5.8. Motif 11 in the chiller power consumption is shown using blue boxplots and the atypical chiller operation is shown using the red solid line. Given the same building load demand, the atypical operation results in much higher chiller power consumption during the period from 15:00 to 19:30. The mean chiller coefficient of performance (COP) decreases from 5.82 to 5.12 (i.e. 12% drop in energy efficiency) when the atypical operation takes place. It is found out by examining original data that during chiller Motif 11, three chillers are running at a nearly full-load condition. By contrast, 4 chillers are switched-on during the atypical operation with a lower part-load ratio. In such a case, the identified atypical operation resulted in a less energy efficient operation.



(a) Antecedent motif (b) Consequent motif

Figure 5.7 Association between building cooling load and chiller motifs in Cluster 4

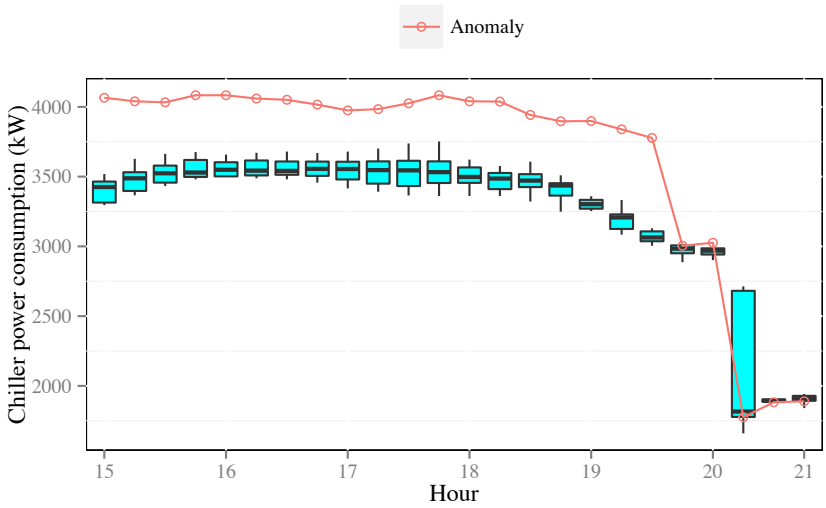
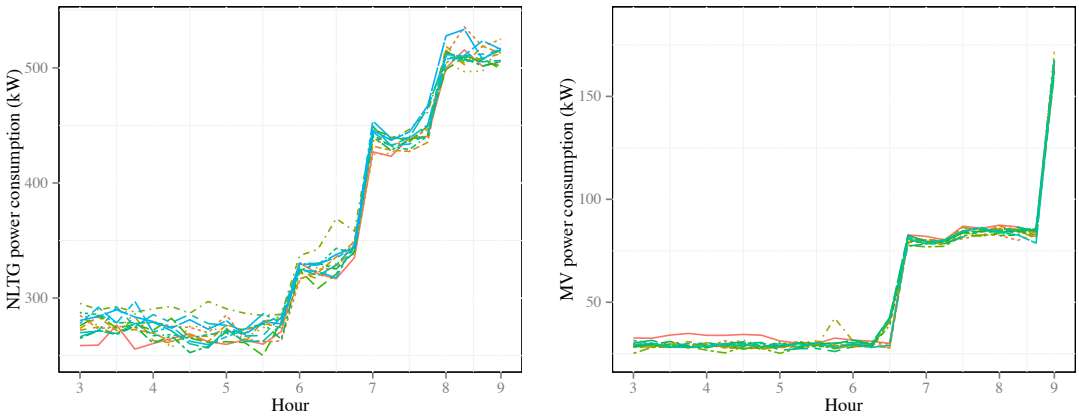


Figure 5.8 Comparison of chiller operations

Another example rule is *NLTG=Motif 6* → *MV=Motif 9*. It describes the association in the operation patterns of the normal power and lighting (NLTG) and mechanical ventilation (MV). These two motifs both take place between 3:00 to 9:00 and are shown in Figure 5.9. Figure 5.10 compares an atypical MV operation with the MV Motif 9. Starting from 4:30, the atypical operation has higher MV consumption than that in MV Motif 9. Further investigation shows that the difference is caused by the MV at the third mechanical floor (i.e., 78/F). Normally, the MV consumption at the third mechanical floor is maintained at around 20kW between 3:00 to 9:00. During the atypical operation, it experiences a sudden increase from 20kW to 45kW at 4:30 and is maintained at that level afterwards. One possible reason for this increase is due to the occupancy change in the corresponding office zone. However, the occupancy in office zone is unlikely to change at 4:30. In addition, the NLTG consumption is also subject to the influence of occupancy and no significant

difference is observed during atypical operation. Such atypical operation may be due to the interference of manual control.



(a) Antecedent motif (b) Consequent motif

Figure 5.9 Association between NLTG and MV motifs in Cluster 4

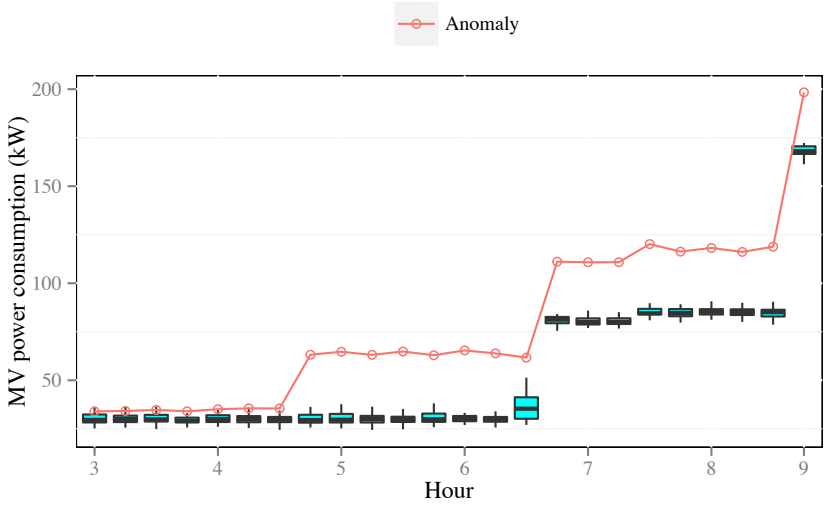
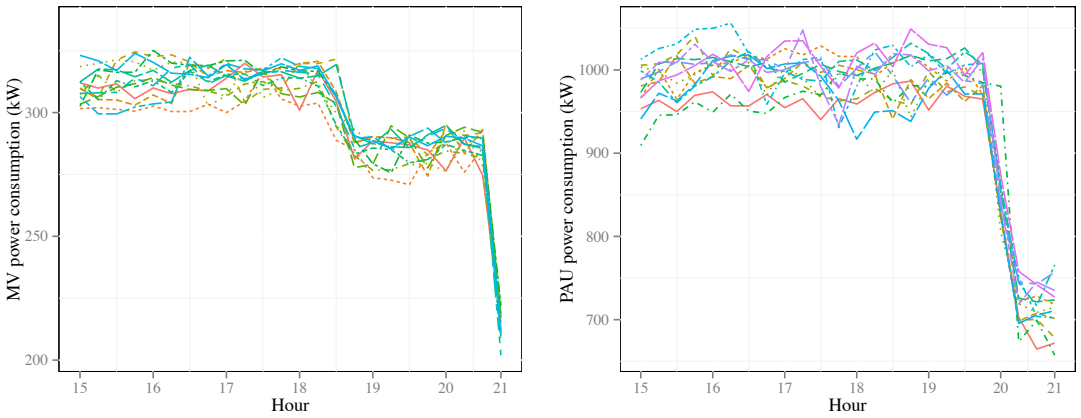


Figure 5.10 Comparison of MV operations

Another example rule describes the association between MV Motif 18 and PAU Motif 14. As shown in Figure 5.11, both motifs take place between 15:00 to 21:00. The two drops in MV consumption at around 18:30 and 20:45 are due to the decrease

in MV consumption at the second and the first mechanical floors respectively. By contrast, one significant drop in the PAU consumption is observed at around 20:00, which is due to the huge decrease in office occupancy. An atypical operation is identified and its PAU consumption is compared with the PAU Motif 14 in Figure 5.12. Compared with PAU Motif 14, the PAU consumption in atypical operation is much smaller from 17:30 to 20:00. The reason behind is that the next day is a public holiday in Hong Kong and many offices have their employees released at around 17:00. Consequently, a power reduction in PAU consumption is observed. In such a case, the atypical operation identified is a normal but rare operation.



(a) Antecedent motif (b) Consequent motif

Figure 5.11 Association between MV and PAU motifs in Cluster 4

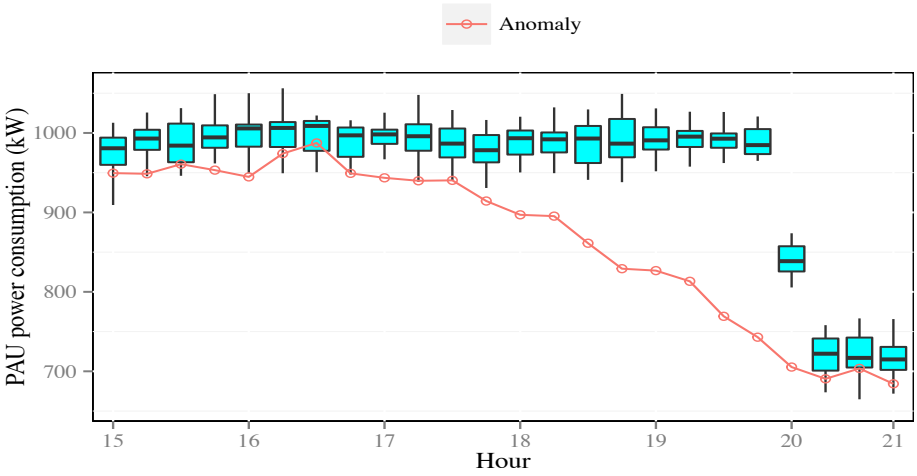


Figure 5.12 Comparison of PAU operations

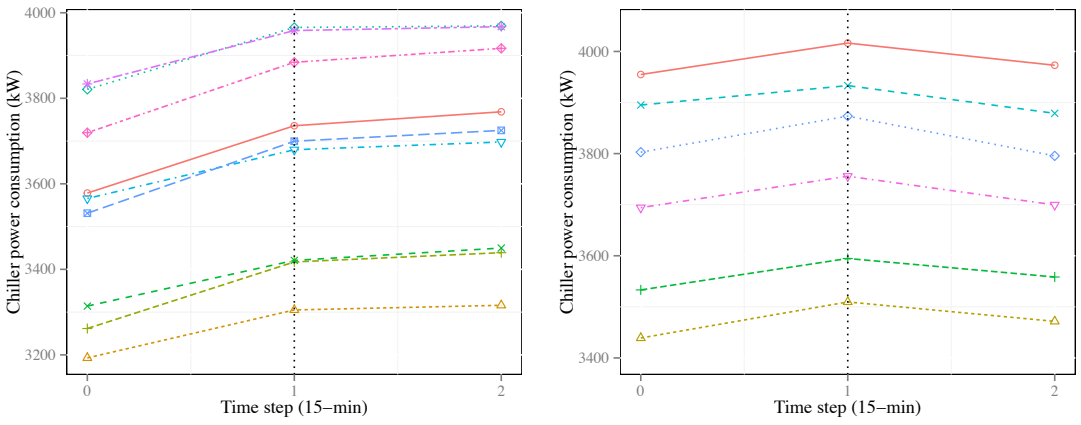
5.3.2 Application of Temporal Association Rules

Temporal Anomaly Detection

Temporal anomaly can be detected using the temporal association rules. Two approaches are possible. If the anomaly widely exists in the time series data, a temporal association rule specifying such atypical association will be derived. In such a case, temporal anomaly can be detected by finding those observations which are in accordance with these temporal association rules. However, those anomaly data are seldom available. The second approach is more practically feasible. A knowledge database of normal temporal association rules can be constructed. Temporal anomaly can be detected by finding those observations which fail to meet the rules in the database.

An example is given here. Two rules with a time lag of 15-minute are derived to

describe temporal associations in the chiller operation between 6:00 to 12:00: *Chiller=High, 5* $\xrightarrow{T=1}$ *Chiller=High, 4* and *Chiller=High, 5* $\xrightarrow{T=1}$ *Chiller=High, 3*. These two rules specify that two possible operation modes are possible at time *T+1* given the chiller power consumption at time *T* is *High* and has a slightly increasing trend. Figure 5.13 presents the subsequences which fulfill these two rules. The chiller power consumption at *T+1* will remain at *High* level, with either a steady or a slightly decreasing trend. Temporal anomalies can be detected by finding subsequences which fail to meet the rule consequent given the same antecedent. Figure 5.14 presents an example of such anomalies. The anomaly is shown in red solid line. It meets the rule antecedent at time *T*; however, the operation mode at time *T+1* becomes *Medium* and has a significant decreasing trend. Further investigation shows that at 8:30, Chiller 4 was switched off while two other chillers were switched on as replacement. After consulting with the operation staff, it is found that Chiller 4 was manually switched off due to its high operation current.



(a) Chiller=High, 5 $\xrightarrow{T=1}$ Chiller=High, 4 (b) Chiller=High, 5 $\xrightarrow{T=1}$ Chiller=High, 3

Figure 5.13 Examples of temporal associations in chiller operation

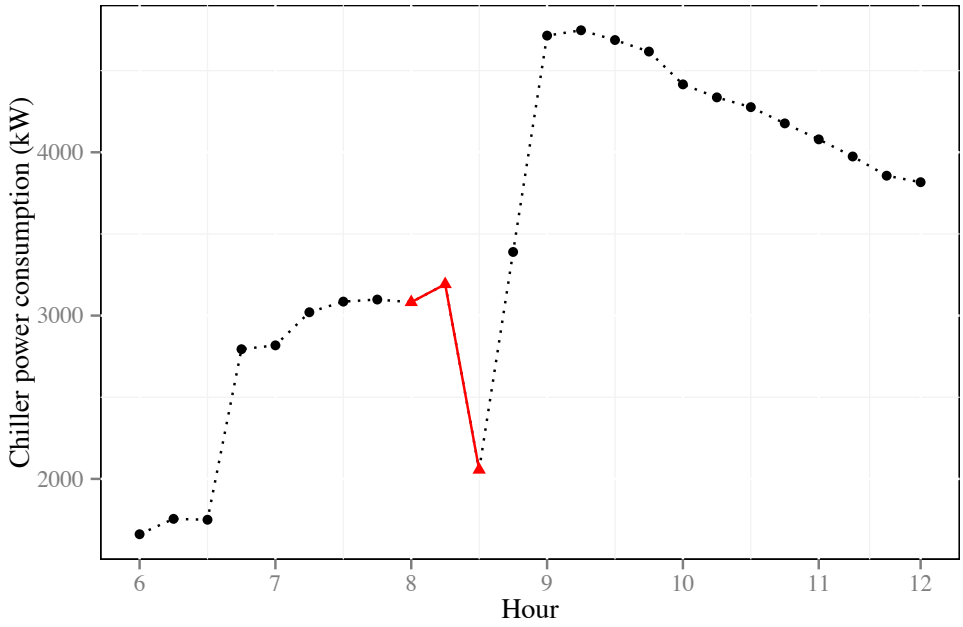


Figure 5.14 An example of temporal anomalies

Characterization of Building Dynamics

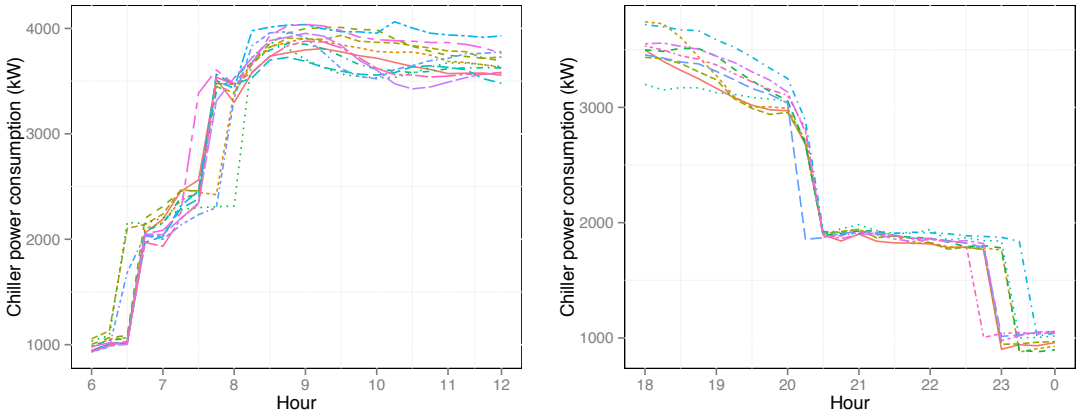
The extraction of time lag between the antecedent and the consequent of temporal association rules helps to characterize the building dynamics. Table 5.5 presents six example rules describing temporal associations in chiller operation. These rules are extracted from two chiller motifs, which are shown in Figure 5.15. The first three rules are derived from the chiller Motif A, which takes place between 6:00 to 12:00. Rule 1 indicates that if the chiller power consumption is *Low* and experiencing a significant increase at time *T*, the operation mode will be at *Medium* level and under slight increase at *T+3*. The second rule shows that once the chiller power consumption starts to increase significantly at *Low* level, it will reach its steady state

at *High* level at *T+10*. Rule 3 shows that the time lag between two significant increases at *Low* and *Medium* levels is around 1 hour. The latter three rules are derived from the chiller Motif B, which occurs between 18:00 to 0:00. Rule 4 states that if the chiller consumption is *High* and experiencing a significant decrease at time *T*, its steady state at *Medium* level will be reached at *T+3*, i.e., 45 minutes later. Similarly, Rule 5 describes that the time needed for the chiller power consumption to reach its steady state from the *Medium* to *Low* level is also 45 minutes. The last rule quantifies that the time lag between the huge decrease at *High* and *Medium* levels is around 3 hours. The result is verified by checking Figure 5.15. The knowledge discovered in this subsection helps to quantify the building dynamics from two perspectives, i.e., the power consumption level and relative changes between successive time steps (i.e., trend). The temporal interactions and dynamics can be automatically extracted. Useful insights can be gained into how building subsystems react to a certain change in operation over time. The temporal associations discovered can be used to facilitate the optimal control and decision-makings in building operation, e.g., chiller sequence control and integration between individual buildings and large power grid systems.

Table 5.5 Temporal associations in chiller operations

Rule	Motif	Antecedent	Consequent	Time lag (15-minute per unit)	Support	Confidence

1	A	Chiller=Low, 7	Chiller=Medium, 5	3	0.92	0.97
2	A	Chiller=Low, 7	Chiller=High, 4	10	0.81	0.92
3	A	Chiller=Low, 7	Chiller=Medium, 7	4	0.83	0.83
4	B	Chiller=High, 1	Chiller=Medium, 4	3	0.36	0.87
5	B	Chiller=Medium, 1	Chiller=Low, 4	3	0.78	0.85
6	B	Chiller=High, 1	Chiller=Medium, 1	12	0.44	0.84



(a) Motif A: Between 6:00 to 12:00 (b) Motif B: Between 18:00 to 0:00

Figure 5.15 Two examples of chiller operation motifs

5.4 Summary

BAS data are in essence multivariate time series data. Currently, few studies have addressed temporal knowledge discovery and applications in big BAS data. This Chapter proposes a generic methodology for mining temporal knowledge from massive BAS data. A diversity of time series data mining techniques and their practical potentials in analyzing big BAS data for building operations and performance management are explored in this Chapter. Rather than addressing pre-defined specific problems, the methodology developed mainly aims to discover

unknown temporal knowledge by adopting unsupervised DM techniques to mine the big BAS data. The intention is to let the data tell the story and then, using domain knowledge to interpret, select and apply the knowledge discovered. The methodology proposed serves as a prototype of big data analysis tools which can be integrated with modern building automation systems to realize automatic knowledge discovery and applications.

This chapter specifically addresses two major challenges in mining big BAS data. One major challenge is the heavy computational load caused by the massive data amount. From a technological perspective, this challenge can be tackled by using high-performance computing machines or cloud-based computing. The adoption of suitable data transformation methods and more computationally efficient DM algorithms can provide an alternative solution. This chapter shows that the SAX method is capable of reducing the data numerocity while preserving the majority of the information contained in the BAS power consumption data. The univariate motif discovery algorithm adopted in this study is based on the concept of combinatorial search rather than exhaustive search and thereby the required computational costs can be largely reduced. Another challenge is the extraction of new features based on the original data for knowledge discovery, also known as feature engineering. Extraction of novel and unique features can greatly enhance the mining result quality. Besides the power consumption level, this chapter makes use of the changing trend to describe the mode of each subsystem at each time step. The temporal association rules

discovered are more meaningful and straightforward for knowledge interpretation and application, compared with those obtained using other features as inputs (e.g., the power consumption level alone).

Time series data mining can discover large amounts of knowledge with different types, such as clusters, univariate and multivariate motifs, and temporal association rules. It is challenging and time-consuming to interpret and apply the knowledge discovered. This chapter develops two methods for the efficient post-processing of knowledge discovered. The first method uses a co-occurrence matrix to map the relationship between univariate motifs. Reliable associations between univariate motifs are derived which provides a novel and convenient approach to utilizing univariate motifs. The second method utilizes a filtering method to improve the temporal association rules mining algorithms with the accurate estimation of time interval between the antecedent and the consequent. The time interval or lag provides valuable insights into building dynamics and HVAC performance characteristics. The methodology has been applied to analyze the BAS data retrieved from the tallest building in Hong Kong. The knowledge discovered has been successfully used to identify anomalies in building operations and characterize the building dynamics. The open-source software *R* and *SPMF* were used to perform the mining.

CHAPTER 6 DEVELOPMENT OF METHODOLOGY FOR THE GRAPH-BASED KNOWLEDGE MINING AND ITS APPLICATIONS

The methodologies proposed in the previous two chapters are designed for mining cross-sectional and temporal knowledge from building operational data with a typical data structure, which uses a single two-dimensional data table to store the data. Actually, the majority of DM techniques are designed to perform the mining task based on such data structure. Nevertheless, the advance in building technologies has imposed new challenges to building professionals, i.e., data are being collected throughout the whole building lifecycle, and are of different types (e.g., text data, video data and numeric data) and structures (e.g., multi-relational databases).

A notable trend in the building field is the implementation of the Building Information Modeling (BIM) technology. BIM intends to provide a digital representation of physical and functional characteristics of a building and has huge potential in evaluating and improving the building lifecycle performance [Schlueter and Thesseling 2009]. It is designed as a structured database to contain all the information about a building from earliest conception to demolition. The data stored in BIM models vary greatly in their data types, such as the schematic drawings showing the spatial information and the text data describing the construction and maintenance projects, and the measurements or control signals of different building

services components in building operations.

An ideal knowledge discovery process would be first effective integrating the information and then performing mining tasks from a unified perspective. Given the complexity in BIM data types and data structures, it is very unlikely that conventional data preprocessing methods (e.g., joining data tables) would fulfill the needs. The current lack in analytic solutions to handle such kind of building data is exactly the challenge that needs to be addressed in the near future. So far, little research has been done to investigate the usefulness of advanced DM techniques on this topic.

This chapter develops a graph-based mining methodology to tackle this problem. Section 6.1 introduces the background of graph-based data mining and the techniques used in this chapter. Section 6.2 presents the research methodology. The methodology is validated through a case study using the BAS data retrieved from the Zero Carbon Building (ZCB) in Hong Kong and is presented in Section 6.3. Section 6.4 summarizes this chapter.

6.1 An Overview of Graph-Based Data Mining

Graph-based DM is the most widely used techniques in analyzing data with complex structures [Cook and Holder 2000]. It has been successfully used to discover useful knowledge in bioinformatics, financial services, counter-terrorism, social network analysis and etc. [Washio and Motoda 2003; Cook and Holder 2006;

Samatova et al. 2013]. Graph is one of the most generic, natural, and interpretable formats for data representation. Great flexibility can be provided in the knowledge discovery process as users can readily manipulate the graph layout to integrate and represent various types of information. In addition, the knowledge discovered using graph-based DM is represented as graphs, which are highly interpretable.

A graph G consists of a set of nodes (or vertices), denoted as $V(G)$ and a set of links (or edges), denoted as $E(G)$. A graph S is said to be subgraph of graph G if $V(S) \subseteq V(G)$ and $E(S) \subseteq E(G)$. A node usually represents an object or a discrete piece of information, while the links are used to represent the relationships between nodes. To illustrate, Table 6.1 presents the power consumption of a chiller and a cooling tower at time T_1 and T_2. Table 6.2 records the spatial information of these two components, one in basement and one on rooftop. It is hard to integrate these two pieces of information into a single two-dimensional data table, as there are three types of information, i.e., temporal, spatial, and power information. By contrast, a graph can be readily constructed for information representation. As shown in Figure 6.1, the graph has 6 nodes and 7 links and they are all labeled. The top 2 nodes are used to represent the temporal information and are labeled as "T_1" and "T_2" respectively. The link connecting these two nodes are labeled as "$dT=1$" which means that the difference in time step is 1. Each of the top 2 nodes is connected with two nodes labeled as "Chiller" and "CT". The labels associated with each edge record the power consumption. The bottom two nodes store the spatial information and are connected

with two components accordingly.

Table 6.1 An example data set containing the power data at two time steps

Time/Power	Chiller	Cooling tower
T1	Low	Low
T2	High	High

Table 6.2 An example data set containing the location of two components

Component	Location
Chiller	Basement
Cooling tower	Rooftop

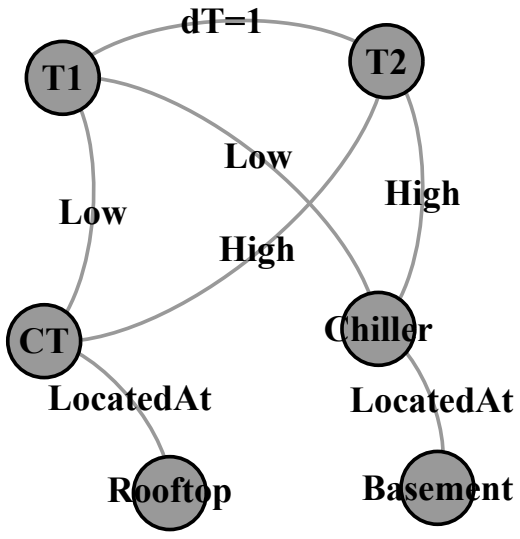


Figure 6.1 An example graph

6.1.1 Proximity Measures for Graphs

Proximity measures are used to evaluate the similarity or dissimilarity of two observations. Many DM tasks, including clustering, classification and anomaly detection, are performed based on data proximity. Conventional data representations usually use a feature vector to describe an observation and the similarity or dissimilarity between two observations can be easily calculated using distance metrics, such as the Euclidean distance.

Proximity measures for graph data can be generally divided into two types, between-graph and within-graph measures [Samatova et al. 2013]. Between-graph measures evaluate the similarity between a set of graphs while with-in graph measures evaluate the similarity between nodes in a graph. One type of between-graph measures transforms a graph to a numeric vector using a set of graph-level indices (GLIs). Conventional measures, such as the Euclidean distance and cosine similarity, can then be used to evaluate the proximity between two graphs. Some commonly used GLIs are introduced as follows. V_G and E_G are the total number of nodes and links in the graph. The degree of a node refers to the number of links associated with it. If the graph is directed, one may further distinguish between in-degree and out-degree, depending on whether the node is used as a head or a tail. The mean degree is the average degree considering all nodes in a graph. The graph density quantifies the ratio between the number of links and the number of possible links and is $\frac{2E_G}{V_G(V_G-1)}$ and $\frac{E_G}{V_G(V_G-1)}$ for undirected and directed graphs respectively.

The graph diameter is defined as the largest distance between any pair of nodes in the graph. The graph transitivity measures the relative frequency of triangles in the graph and is defined as $\frac{3\times No.of\ triangles}{No.of\ connected\ triplets}$. Such type of methods is easy to implement. However, the GLIs can only capture the topological information of a graph. It is generally not applicable to labeled graphs.

The other type of between-graph measures is the edit distance [Cook and Holder 2006]. The edit distance between two graphs refers to the smallest cost resulted from a set of edit operations to transform one graph to another. Typical edit operations include the insertion, deletion and substitution of nodes and links. It can be applied to capture the proximity between labeled graphs. The costs associated with different edit operations can be either user-defined or optimized through a training process [Gao et al. 2010].

6.1.2 Frequent Subgraph Mining

Frequent subgraph mining (FSM) is regarded as the essence of graph-based data mining [Jiang et al. 2004]. FSM mainly works on undirected graphs with labeled nodes and links. Popular applications of FSM include finding the common substructures of chemicals and identify the frequent patterns of terrorist attacks [Jiang et al. 2004; Samatova et al. 2013].

In this section, some representative FSM algorithms which take a set of graphs

as input are introduced. These algorithms can be classified based on two criteria, i.e., whether the search is exact or inexact, and whether the search strategy is breadth-first or depth-first [Jiang et al. 2004]. Inexact search FSM algorithms, such as SUBDUE [Cook and Holder 1994] and CREW [Kuramochi and Karypis 2004], use approximated measures to compare two graphs. The resulting mining efficiency is high. However, it is not guaranteed to discover all frequent subgraphs. Exact FSM algorithms are more commonly used due to their ability of discovering all frequent subgraphs. Some adopt the breadth-first search (BFS) strategy to generate subgraph candidates. The basic concept is that a subgraph with a node size of *(k+1)* cannot be frequent if any of its parent subgraphs of a node size of *k* is not frequent. Popular algorithms belonging to this category include AGM [Inokuchi et al. 2000] and DPMine [Vanetik 2002]. The BFS strategy can greatly reduce the number of redundant candidates generated. However, creating candidate with a size of *(k+1)* based on frequent subgraphs with a size of *k* can be computationally expensive, especially when *k* is large. Therefore, the BFS strategy usually suffers from a problem of poor computer memory utilization [Jiang et al. 2004]. The depth-first search (DFS) can better utilize the computer memory when generating subgraph candidates (e.g., through the right-most extension [Yan and Han 2002]). In addition, some procedures have been developed to make it more efficient to test whether two subgraphs are identical or not. As a result, the DFS-based algorithms have become the main approach to FSM. Some representative algorithms in this category include MoFa

[Borgelt and Berthold 2002], gSpan [Yan and Han 2002], FFSM [Huan et al. 2003] and GASTON [Nijssen and Kok 2004]. A recent study compared the performance of these four exact DFS-based FSM algorithms [Worlein et al. 2005]. The gSpan algorithm generally has better performance considering the running time and memory usage.

One essential challenge of FSM is that the number of frequent subgraphs discovered can be very large and the majority of them are redundant. A subgraph becomes redundant if there is a supergraph which has the same support count. To enhance the mining efficiency, Yan and Han proposed an algorithm called CloseGraph to mine closed frequent graphs based on their work of gSpan [Yan and Han 2003]. A subgraph is called closed if there exists no supergraph having the same support count. It was shown that CloseGraph could dramatically reduce the number of redundant subgraphs and therefore, enhancing mining efficiency. In this study, the CloseGraph is adopted to mine frequent subgraphs. The algorithm takes a set of graphs as input. Users need to define a minimal support threshold, which is used to evaluate whether a subgraph is frequent or not.

6.2 Graph-based Knowledge Discovery Methodology

This chapter develops a graph-based data mining methodology to mine BAS data with potentially complex structures. The methodology is developed based on the

generic framework proposed in Chapter 3. The outline is shown in Figure 6.2. At the data exploration phase, two types of graph generation methods, i.e., observation-based and variable-based methods, are proposed to transform the BAS data into graphs. The decision tree method is applied for data partitioning. Frequent subgraphs are discovered using the CloseGraph algorithm at the knowledge discovery phase. Two post-mining methods are proposed to enhance the efficiency and effectiveness in knowledge interpretation, selection and application.

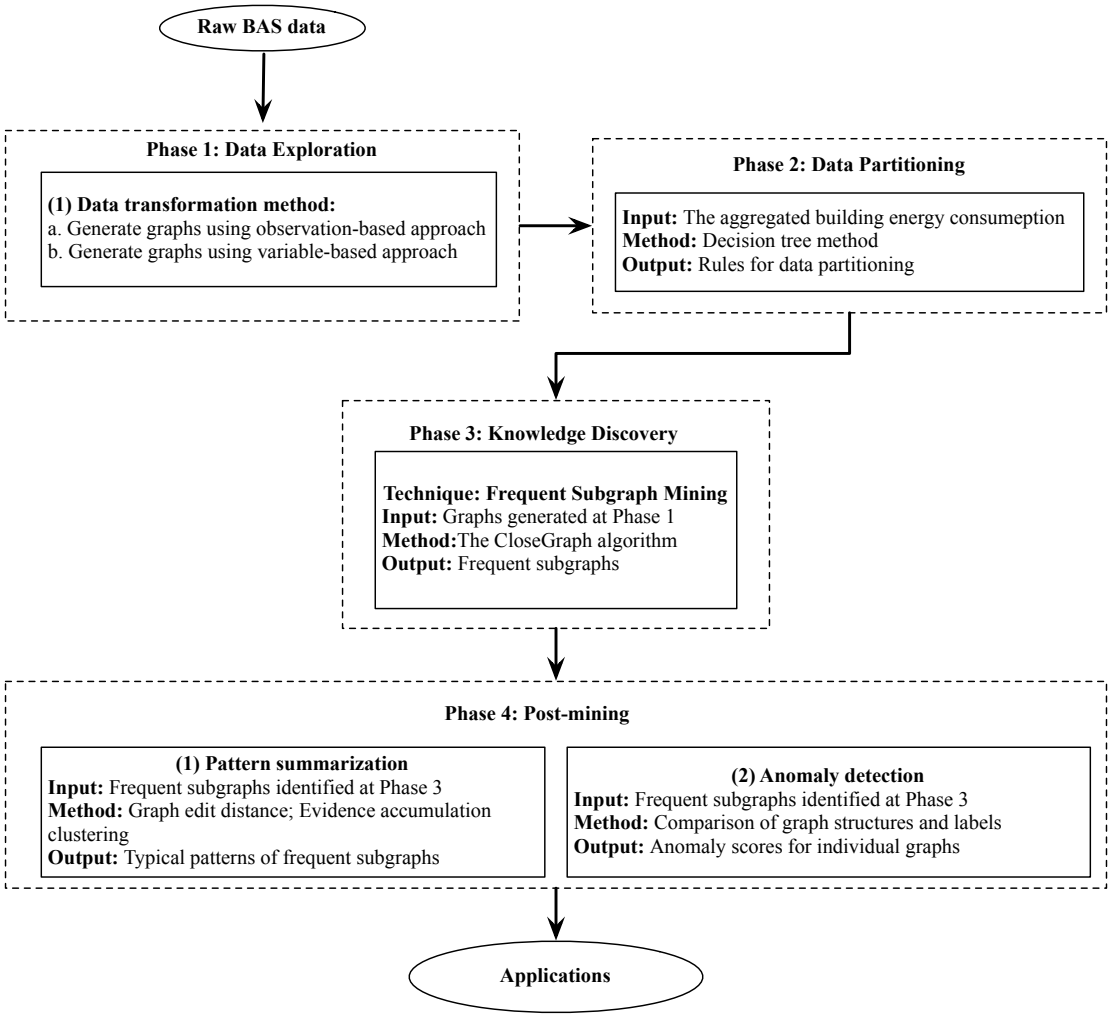


Figure 6.2 Research outline of graph-based data mining for BAS data

6.2.1 Data Exploration

Apart from the data cleaning methods introduced in previous two chapters, the data exploration phase of this methodology specifically addresses the graph generation problem, i.e., generate graphs based on BAS data. The basic format of BAS data is a two-dimensional data table, where each row represents an observation recorded at a time step and each column represents a numeric or categorical variable. Two approaches are developed to transform BAS data into graphs, namely the observation-based and variable-based approaches. Both approaches are developed taking into account the following key considerations, i.e., the computation efficiency and the compatibility with FSM algorithms. To ensure the computation efficiency, the graph should be created in such a manner that the number of nodes and links used to describe a certain amount of information is minimum. The second consideration requires the graph to be generated is unweighted, labeled and connected (i.e., there is always a path from one node to another). Discretization should be performed to transform numeric variables into categorical variables. The details of these two approaches are introduced as follows.

The Observation-based Approach

The observation-based approach provides a straightforward way to transform the raw BAS data into graphs. In this study, it is designed to represent three types of

information in the BAS data, i.e., the temporal information, the level and trend information. The level information refers to the value of a variable at time t (i.e., denoted as y_t). The trend information refers to the relative change between successive time steps (i.e., denoted as y_t-y_{t-1}). The graph to be generated contains two parts. The first part includes the nodes representing the temporal information, i.e., temporal nodes. Links between temporal nodes are established to represent the time flow. The second part consists of the nodes representing the level information of different variables, i.e., level nodes. Links are established between temporal nodes and level nodes. The trend information is encoded into the graph by adding edge labels between temporal and level nodes. Assuming that the BAS data have N observations recorded in a chronological order and p variables, the first and second parts will contain N and M nodes respectively, where $M = \sum_{i=1}^{p} |L_i|$ and $|L_i|$ is the number of possible values for the i^{th} categorical variable. The graph to be generated has $(N+M)$ nodes and $(p+1)\times N-1$ links.

Table 6.3 The temporal and level information of three variables

Time/Level	Chiller	Cooling tower	AHU
T_1	Low	Low	High
T_2	High	High	High
T_3	High	High	High
T_4	High	High	Low
T_5	Low	Low	Low

T_6	Low	Low	Low

As an example, Tables 6.3 and 6.4 present the power consumption level and trend information of three variables at 6 consecutive time steps, i.e., T_1 to T_6. The graph generated using the observation-based approach is shown in Figure 6.3. The temporal and level nodes are shown as grey and green circles respectively. The first part of the graph consists of 6 temporal nodes and links are established to represent the time flow. Each of the three variables contains 2 possible levels, i.e., *Low* and *High*. Therefore, the second part contains $M = \sum_{i=1}^{p} |L_i| = 2 + 2 + 2 = 6$ level nodes. The links between temporal nodes and level nodes are labeled according to the trend information.

Table 6.4 The temporal and trend information of three variables

Time/Trend	Chiller	Cooling tower	AHU
T_1	Steady	Steady	Increase
T_2	Increase	Increase	Steady
T_3	Steady	Steady	Steady
T_4	Steady	Steady	Decrease
T_5	Decrease	Decrease	Steady
T_6	Steady	Steady	Steady

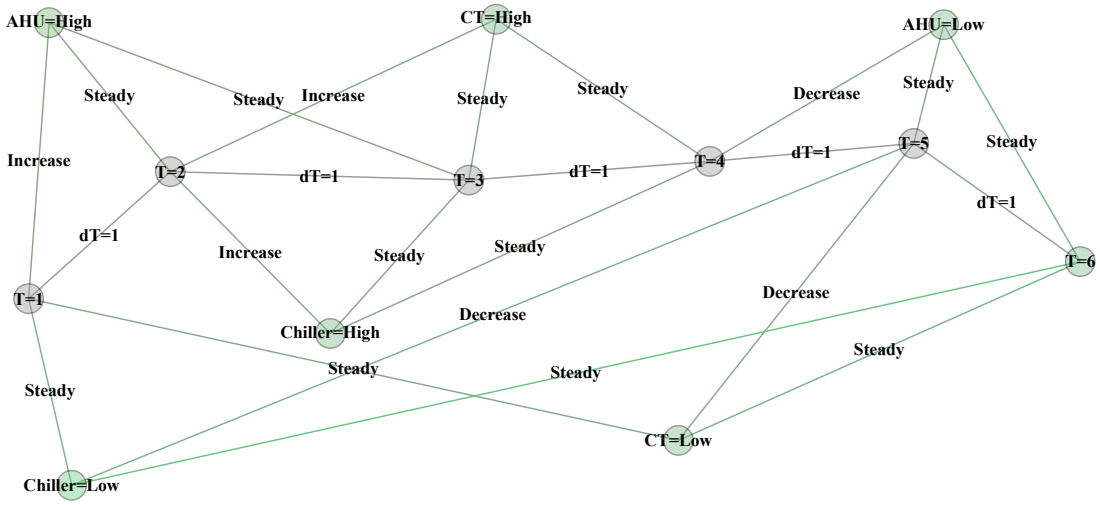


Figure 6.3 The example graph generated by the observation-based approach

The Variable-based Approach

The variable-based approach developed is inspired by the graph representation used in social network analysis, where each individual is represented as a node and their associated relationships are shown as links. The general idea is that each BAS variable is in analogy to an individual person and denoted as a node. The interactions between BAS variables during a certain time period are modeled as links. The observation-based approach is capable of preserving the detailed or low-level information in the original BAS data. By contrast, the variable-based approach focuses on extracting the abstract or high-level information in the original BAS data.

The main challenge of the variable-based approach is to come up with a meaningful way to describe the interaction between two variables. The resulting graphs should be able to provide meaningful and useful insights once their frequent subgraphs are discovered in the later step. An intuitive type of methods to describe the

interaction between two variables during a certain time period is to calculate their correlation. This type of methods are easy to implement and can work with both numeric and categorical variables in the BAS data. However, the information conveyed in the resulting graph can be too abstract to provide any insightful knowledge in the later mining process. For instance, a high correlation between two numeric variables does not provide any indication on the actual operation conditions (e.g., whether the power consumption is at a low or high level), which are usually the main concerns in building management.

This research proposes a novel method to create link labels to represent the interaction between two variables in the BAS data. The method works with categorical variables and therefore, discretization should be performed for numeric variables. Assuming that the BAS data to be transformed to a variable-based graph has N observations, the first step is to determine a window size (denoted as w), which is used to divide the BAS data into $\frac{N}{w}$ non-overlapping temporal segments. The dominant, or the most frequent interaction modes between two variables in these temporal segments can be discovered. The interaction mode is defined as a vector containing the categorical values of both variables. A notation is created based on the dominant interaction mode between two variables during each temporal segment. Table 6.5 presents an example of such notation assuming both variables have two levels (denoted as "Low" and "High"). It should be mentioned that there could be a tie in the dominant interaction mode. In such a case, a longer notation is created with

each ends surrounded by zeros, e.g., denoted as "0120" when {*Low*, *Low*} and {*Low*, *High*} are tied as the dominant interaction mode. The link label between these two variables can be obtained by combining the notations in different temporal segments.

As shown in Figure 6.4, an example graph is created using the above-mentioned method for the data shown in Table 6.3. Considering w=2, the whole data will be divided into 3 segments. All three variables are categorical with 2 possible values. Taking the chiller and cooling tower as an example, there is a tie in the dominant interaction mode in the first temporal segment, i.e., {*Low*, *Low*} and {*High*, *High*}. Consequently, the notation for this segment is "0140". The notations for the other two segments are "4" and "1" respectively. The link label between these two variables are "014041".

Table 6.5 Notations for different dominant interaction modes

Variable A	Variable B	Interaction mode	Notation
Low	Low	{Low, Low}	1
Low	High	{Low, High}	2
High	Low	{High, Low}	3
High	High	{High, High}	4

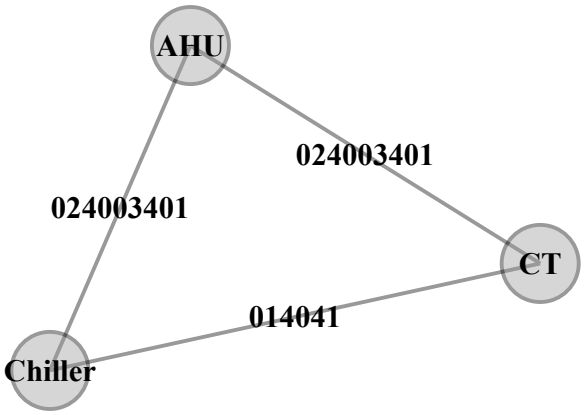


Figure 6.4 An example graph using the variable-based approach

6.2.2 Data Partitioning

Previous two chapters mainly adopt clustering analysis to perform the data partitioning task. Clustering analysis is a natural fit for data partitioning. However, the only output obtained from clustering analysis is the cluster membership of each observation. Since clustering analysis is essentially an unsupervised learning method, it cannot produce straightforward knowledge to guide the data partitioning process. In other words, it is not easy to summarize the characteristics of each cluster. In this chapter, the usefulness of decision tree method for data partitioning is explored. As a supervised learning method, the decision tree model has the ability to capture complex relationships between independent and dependent variables. The model is highly interpretable and the rules derived by the decision tree model can be used as guidance for data partitioning.

More specifically, a decision model is developed to capture and visualize the relationship between the aggregated building power consumption and the time

variables (i.e., “Month”, “Day”, “Day Type”, “Hour”, “Minute”). “Day Type” here refers to Monday to Sunday. It is worth mentioning that the time variables are categorical and have different numbers of possible values (e.g., “Month” has 12 values and “Day Type” has 7 values). It is known that conventional decision tree algorithms (e.g., CART) have a selection bias towards input variables with many possible values [Horthon et al. 2006]. To eliminate such selection bias, the unconditional inference tree method is adopted [Horthon et al. 2006]. The whole BAS data are partitioned based on the relationship discovered by the decision tree model.

6.2.3 Graph-based Knowledge Discovery

The CloseGraph algorithm is applied to discover frequent subgraphs [Yan and Han 2003]. It is in analogy to the discovery of frequent item sets from cross-sectional data. The input is the graphs generated at Phase 1. The algorithm is applied separately to the data partitions identified at Phase 2. The outputs of CloseGraph are the frequent subgraphs among the graph databases. The CloseGraph algorithm from the ParSeMis project is used in this research [Philippsen et al. 2008].

6.2.4 Post-mining

Even though the CloseGraph can greatly reduce the number of redundant subgraphs, the number of frequent subgraphs could still be too large for manual

inspection. Two post-mining methods are developed to improve the efficiency in the post-mining step and their details are explained as follows.

Pattern Summarization

The pattern summarization aims to derive a small set of representative patterns based on the potentially large number of frequent subgraphs discovered. The general idea is to perform clustering analysis on the frequent subgraphs discovered and then use the cluster centroids as the representative patterns.

To perform the clustering analysis, a proximity measure should be adopted to evaluate the similarity between graphs. The graph edit distance is adopted and the costs associated with different edit operations (i.e., insertion, deletion and revision) are assumed to be equal. As a result, a distance matrix can be constructed for the graph data. Then, an ensemble clustering method, the evidence accumulation clustering (EAC), is adopted to perform the cluster analysis. EAC has the ability to discover clusters with various sizes and shapes. In addition, there is no need to explicitly define the cluster number which is usually impossible to know in prior. Users only need to specify the lower and upper limits of the cluster number and the EAC method can automatically determine the optimal cluster number. The partitioning around medoids (PAM) algorithm is selected as the base clustering algorithm. PAM shares a similar procedure with the well-known *k*-means algorithm, but is more robust to noise and outliers and more importantly, it can take distance

matrix as input. The EAC process is iterated with a varying cluster number, which is randomly selected between the lower and upper limits of the cluster number. A collision matrix, denoted as C, is formed to record the percentage of times that two observations are grouped in the same cluster. The hierarchical clustering algorithm is then applied to obtain the final clustering result. We direct interested readers to [42] for more details on EAC.

Anomaly Detection Based on Frequent Subgraphs Discovered

A method is proposed to detect anomalies in the form of graphs. Assuming that Y frequent subgraphs are discovered based on X graphs, the method outputs an anomaly score for each of the X graphs. The general idea is that a graph is abnormal if it has no subgraphs that perfectly match any of the frequent subgraphs discovered. For a given graph G_i, the anomaly score is defined as $A_i = \frac{1}{Y}\sum_{j=1}^{Y}\frac{D_{i,j}}{N_{s,j}}$, where $D_{i,j}$ is the minimal number of differences in nodes and links between any subgraphs of G_i and the j^{th} frequent subgraph, $N_{s,j}$ is the number of nodes and links of the j^{th} frequent subgraph. If there exists a perfect match between any subgraphs of G_i and a frequent subgraph discovered, A_i is assigned as infinity. A larger A_i indicates that G_i is less close to any of the frequent subgraphs discovered. In such a case, G_i may represent some unique and infrequent operation conditions. By contrast, the closer the A_i approaches to zero, the more interesting or potentially useful the anomaly could be. In such a case, it indicates a well-disguised anomaly, i.e., the graphs only differ on a minor scale.

Therefore, it is recommended to manually inspect the graphs according to an increasing order of their anomaly scores.

As an example, Figures 6.5 and 6.6 present a graph generated in the observation-based approach and an example frequent subgraph used for anomaly detection respectively. It is observed that the difference lies in the operation condition of cooling tower at *T=2* and the number of difference *D=2*, i.e., changing the node label from "*CT=Low*" to "*CT=High*" and the link label from "*Steady*" to "*Increase*". The total number of nodes and links of the frequent subgraph is 9 and therefore, the anomaly score is $\frac{2}{9}$.

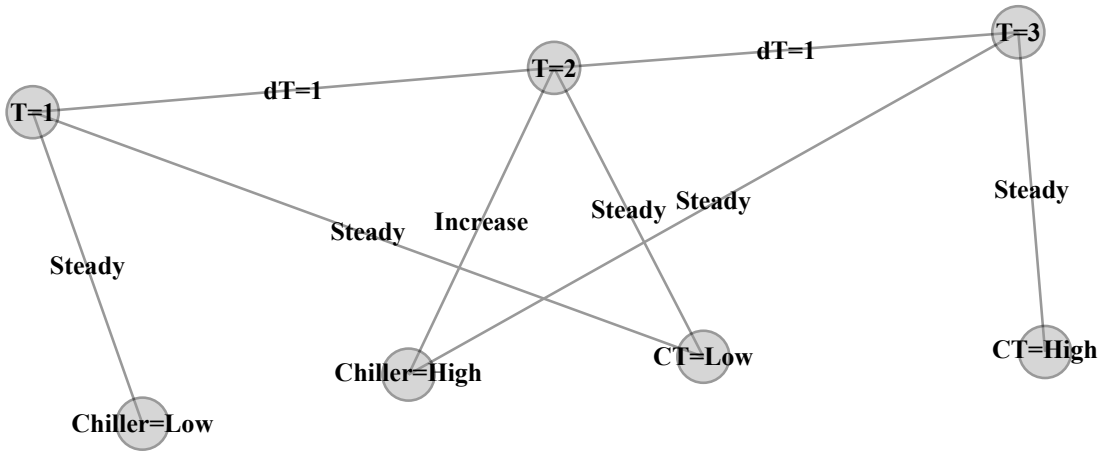


Figure 6.5 An example graph considered for anomaly detection

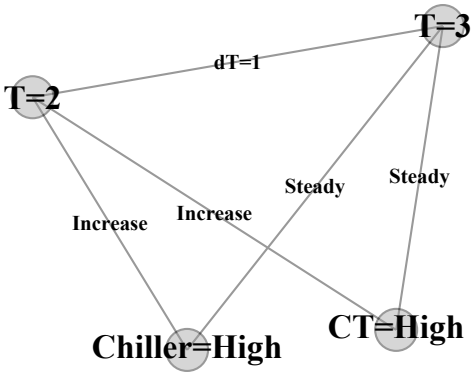


Figure 6.6 An example frequent subgraph discovered for anomaly detection

6.3 Mining Real BAS Data

6.3.1 Data Partitioning Using The Decision Tree Method

The method introduced in section 6.2.2 is used to discover the relationships between the aggregated building power consumption and the time variables, i.e., "Year", "Month", "Day", "Hour", "Day Type". The decision tree developed is shown in Figure 6.7. Three variables, i.e., "Day Type", "Hour" and "Month" are selected as the splitting variables. The root node selects the "Day Type" as the splitting variable. It is found out that ZCB is close on Wednesdays and Sundays. Therefore, the aggregated building power consumption would be lower on Wednesday and Sundays than that in other day types. Nodes 2 and 5 both select the "Hour" as the splitting variable and the splitting criteria coincide with the office hours (i.e., 9:00 to 18:00) and non-office hours of ZCB. Node 7 is added to further partition the building power consumption during the office hours on working days. It selects the "Month" as splitting variable and the splitting criteria matches the seasonality in Hong Kong, i.e., May to October as the hot season and the rest as the cold season.

The knowledge discovered in this step can be used in two ways to guide further mining activities. Firstly, it can be used to partition the data to enhance the sensitivity and reliability of the knowledge discovered in the following steps. Secondly, it helps to provide guidance on mining specific type of knowledge. For instance, domain

knowledge tells us that the building services systems will be unsteady during the transitions between non-office hours and office hours. Therefore, the data recorded during these intervals should be utilized to derive knowledge on the dynamics and characteristics in system operations.

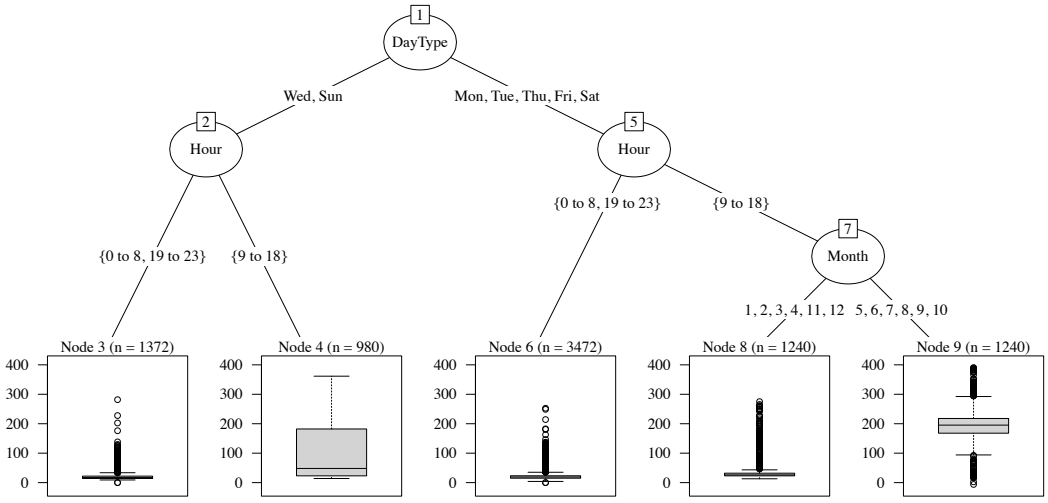


Figure 6.7 The decision tree model developed for the ZCB data

6.3.2 Discovering Representative Patterns in System Operations

The observation-based graphs are well suited to represent the temporal, level and trending information of BAS data. The knowledge discovered from such type of graphs can be valuable for characterizing the patterns and dynamics in system operations. To illustrate, this section investigates the dynamics of HVAC system during the stage-up process. The key variables in the HVAC system, including the load demand, power consumptions of water-cooled chillers (WCC), cooling towers

(CT), air-handling units (AHU) and primary air-handling units (PAU), are transformed into observation-based graphs. The data recorded during 6:00 to 12:00 in the working days of hot seasons are used for knowledge discovery. All variables are discretized into three levels, i.e., "*Idle*", "*Low*" and "*High*" using the *k*-means clustering. The trend data of both variables are discretized into five categories, i.e., "*Huge Decrease*", "*Slight Decrease*", "*Steady*", "*Slight Increase*", "*Huge Increase*". In total, 124 graphs are generated. Each graph has 22 nodes and 41 links. The minimum support threshold used for frequent subgraph mining is 20%. The post-mining method introduced in section 6.2.4.1 is used for pattern summarization. To perform the EAC, the lower and upper limits of cluster number are set as 2 and 20 respectively and the iteration number is fixed as 50. As a result, 298 frequent subgraphs are discovered, based on which 13 representative graphs are summarized. The knowledge obtained is highly interpretable. It is straightforward to understand the temporal dynamics and system operation characteristics.

Two representative patterns are shown in Figures 6.8 and 6.9. In both figures, the PAU stays idle during the whole period and the HVAC system stays idle between 6:00 and 7:00. As shown in Figure 6.8, the Load Demand experiences a huge increase and reaches its "*Low*" level at 8:00. To cope with this change, both WCC and AHU reach their "*Low*" levels with a slightly increasing trend. Meanwhile, the CT stays idle. At 9:00, the Load Demand and AHU continue to increase with a slightly increasing trend and reach their "*High*" levels. The WCC also reaches its "*High*" level but with a

huge increasing trend. The power consumption of CT reaches its "*Low*" level with a huge increase trend. At 10:00, the Load Demand, WCC and AHU stay steady. A slight increase trend is observed in CT as it reaches to its "High" level. All variables are maintained steady after 10:00. Compared to Figure 6.8, the main difference in Figure 6.9 is that none of these variables reach their "High" level. In addition, a slight decreasing trend in WCC consumption is observed at 10:00 before reaching the steady state. From these two figures, it can be inferred that the transient changes in the HVAC stage-up process typically last for 2-hour, i.e., between 8:00 and 10:00.

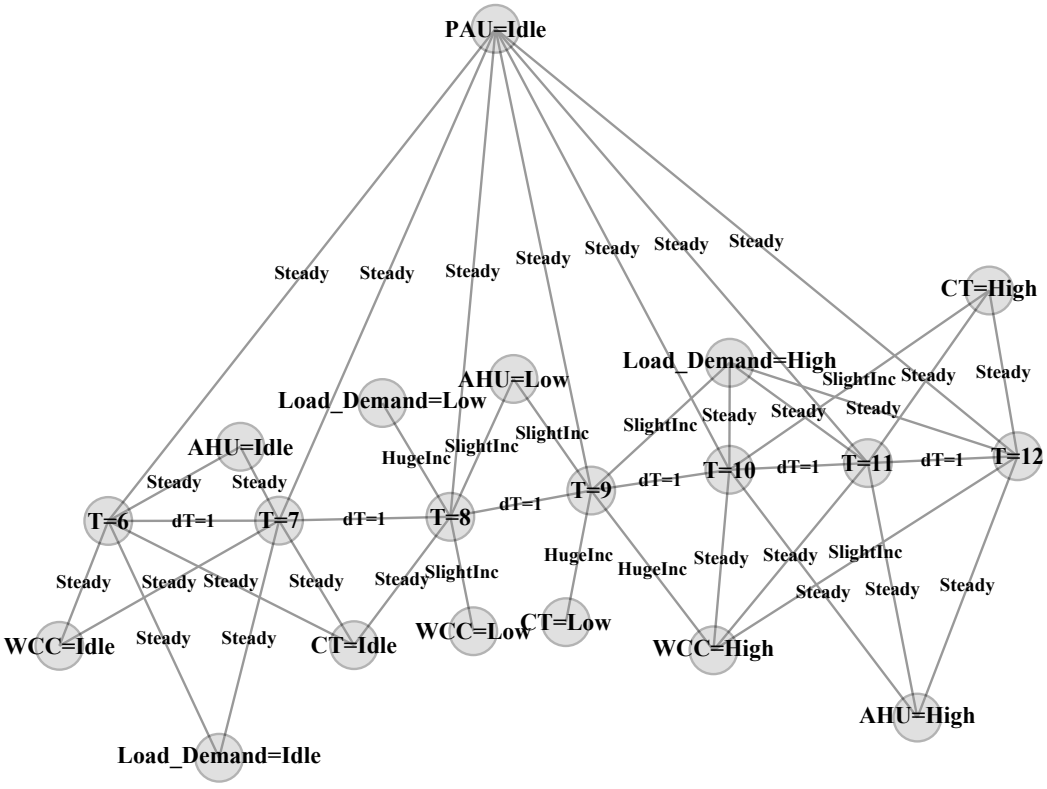


Figure 6.8 Representative graph A

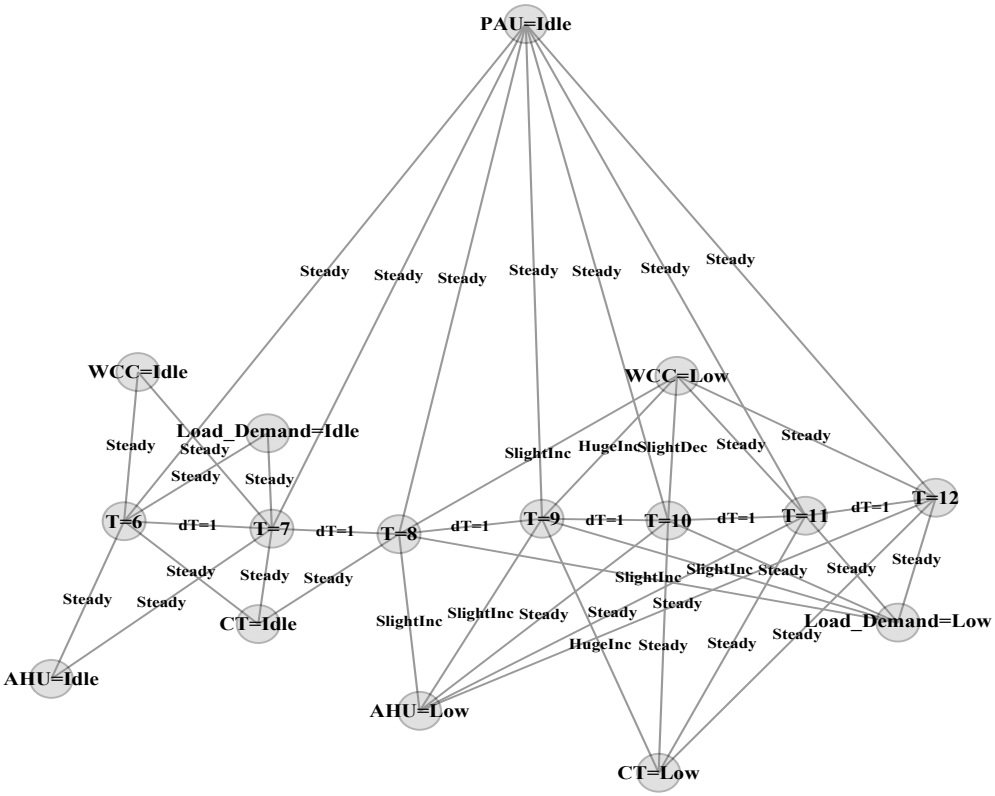


Figure 6.9 Representative graph B

6.3.3 Discovering Atypical Operations

The variable-based graphs are used as high-level abstracts of the BAS data. In this section, the BAS data during the office hours (i.e., 9:00 to 18:00) in the working days of hot seasons are transformed into variable-based graphs with structural, temporal and level information embedded. Each numeric variable is discretized into 3 levels, denoted as "*Idle*", "*Low*" and "*High*". Figure 6.10 presents an example graph generated using the BAS data during office hours on July 4, 2013 (Thursday). The Load Demand is designed as the central node and connected with seven nodes representing the HVAC subsystems, i.e., WCC, AHU, CT, CDWP, CHWP, PAU and BDG. Some subsystems contain multiple components and such structural information

is recorded by establishing links between subsystems and their individual components. The normal power and lighting consumptions at different locations in ZCB are also recorded in the graph. The link labels are created to summarize the interactions between two variables in three temporal segments, i.e., 9:00 to 11:00, 12:00 to 15:00 and 16:00 to 18:00. For instance, the link label between CT and CT 3 is "699". It means that the dominant modes are "*CT=Low, CT 3=High*", "*CT=High, CT 3=High*" and "*CT=High, CT 3=High*" in each temporal segment respectively.

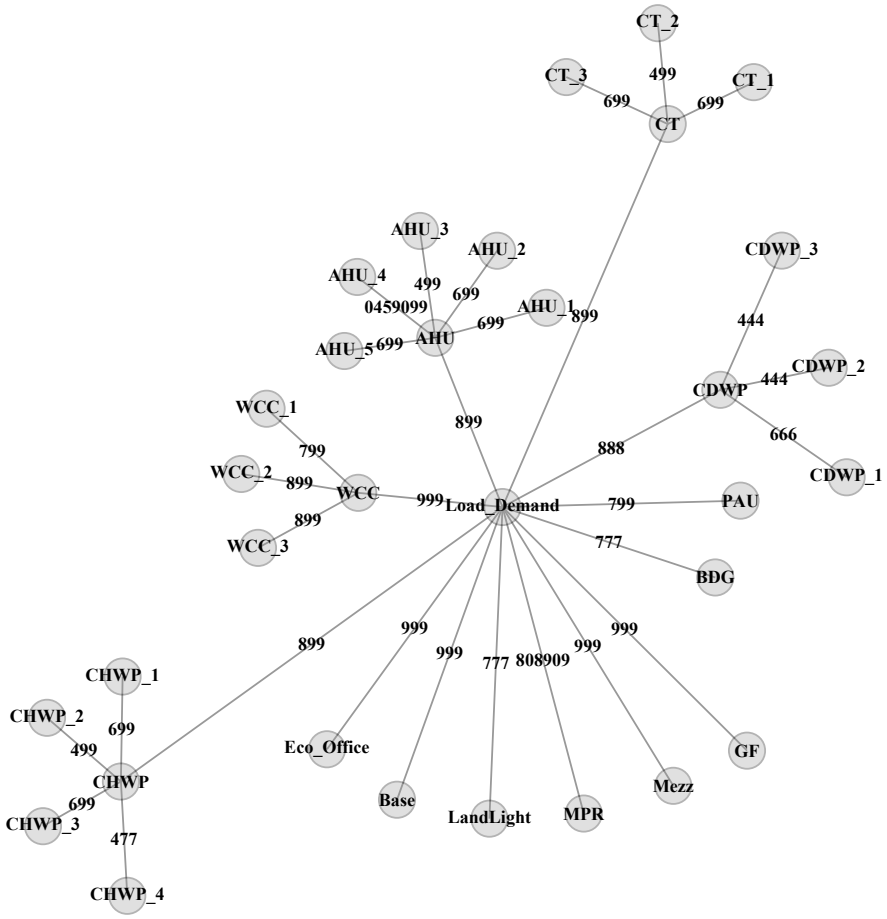


Figure 6.10 An example variable-based graph during office hours on July 4, 2013 (Thursday)

The minimum support threshold for frequent subgraph mining is set as 10%,

which is in accordance with the common definition of anomalies [Lazarevic et al. 2004]. In total, 1082 frequent subgraphs are discovered and used as a knowledge database. The post-mining method proposed in Section 6.2.4.2 are used to find atypical operations. As an example, a graph is identified as atypical with a score of 0.51, indicating that on average, it is different from all the frequent subgraphs discovered with a mean proportion of 51%. Figure 6.11 shows the graph with reference to its closest frequent subgraph. It is created in such a manner that the matched and unmatched portions are shown in blue and pink respectively, and the rest is shown in grey. The frequent subgraph considered is shown in Figure 6.12. It is apparent that the main difference is the Load Demand in the third temporal segment (i.e., 16:00 to 18:00), which is "*High*" in the frequent subgraph and "*Low*" in the atypical operation. Further inspection reveals that the atypical graph represents the building operation during office hours on September 20, 2013 (Friday), which is a public holiday in Hong Kong. After consulting with the operation staff, it is found out that on normal working days, ZCB are open for indoor tours during three time slots, i.e., 10:00 to 11:30, 14:00 to 15:30 and 16:00 to 17:30. However, the last tour does not exist on Wednesdays and public holidays. The resulting load demand during that time period will be smaller than usual. The atypical operation identified is an infrequent but normal operation.

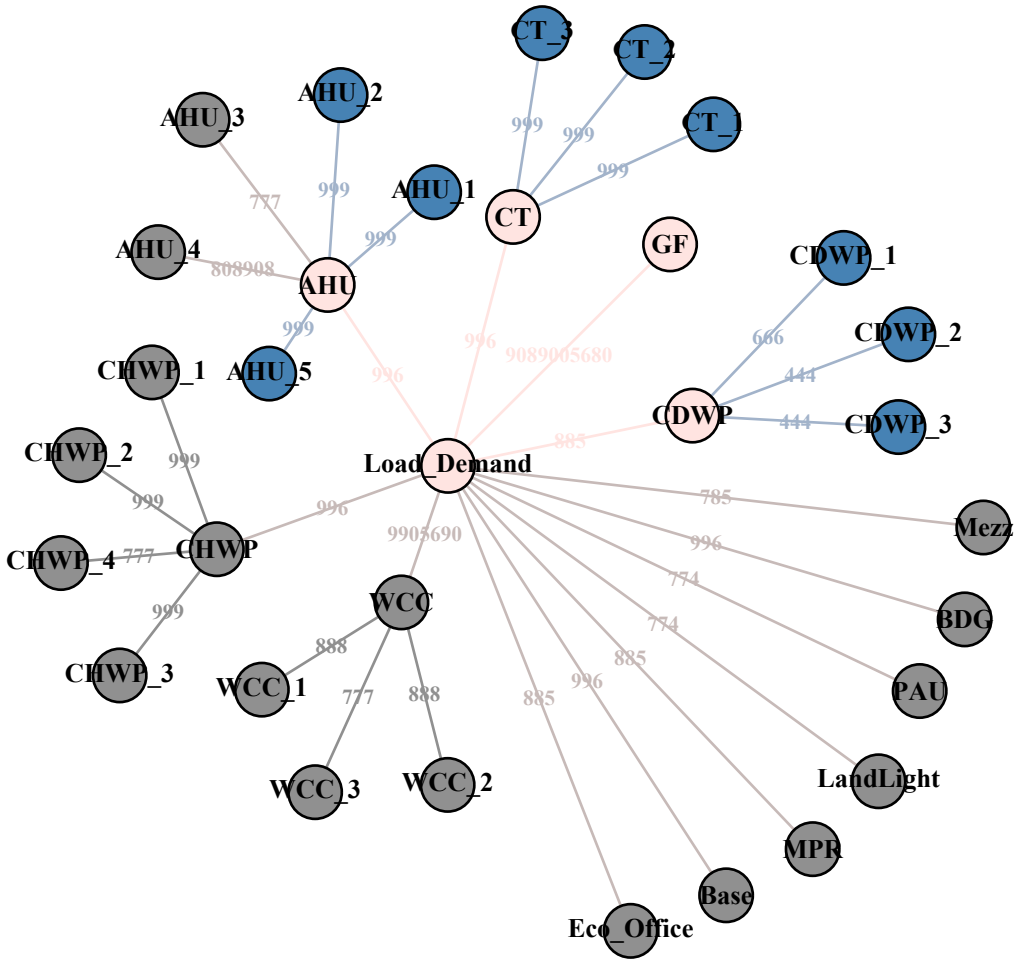


Figure 6.11 An atypical operation on September 20, 2013 (Friday)

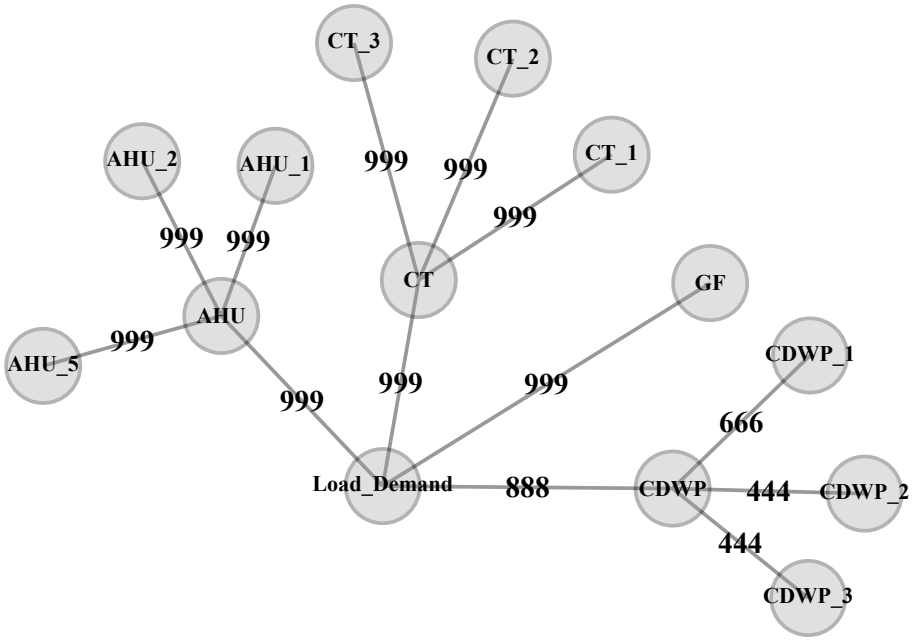


Figure 6.12 The frequent subgraph considered in Figure 6.11

Another atypical operation is identified on July 2, 2013 (Tuesday). The atypical graph is shown in Figure 6.13 with reference to its closest frequent subgraph (shown in Figure 6.14). The link label between "Load Demand" and "PAU" is "799" in the atypical graph, indicating that the dominant interaction modes are "Load Demand=High, PAU=Idle", "Load Demand=High, PAU=High" and "Load Demand=High, PAU=High" during 9:00 to 11:00, 12:00 to 15:00 and 16:00 to 18:00. respectively. By contrast, the dominant interaction modes between these two variables in the frequent subgraph are all "Load Demand=High, PAU=Idle". The PAU system in ZCB rarely operates, as the indoor occupant number in ZCB is usually small and the natural ventilation system can adequately fulfill the fresh air demand. It turns out that ZCB hold a special event on July 2, 2013 and received a large number of visitors. The PAU system was switched on to meet the unsatisfied fresh air demand.

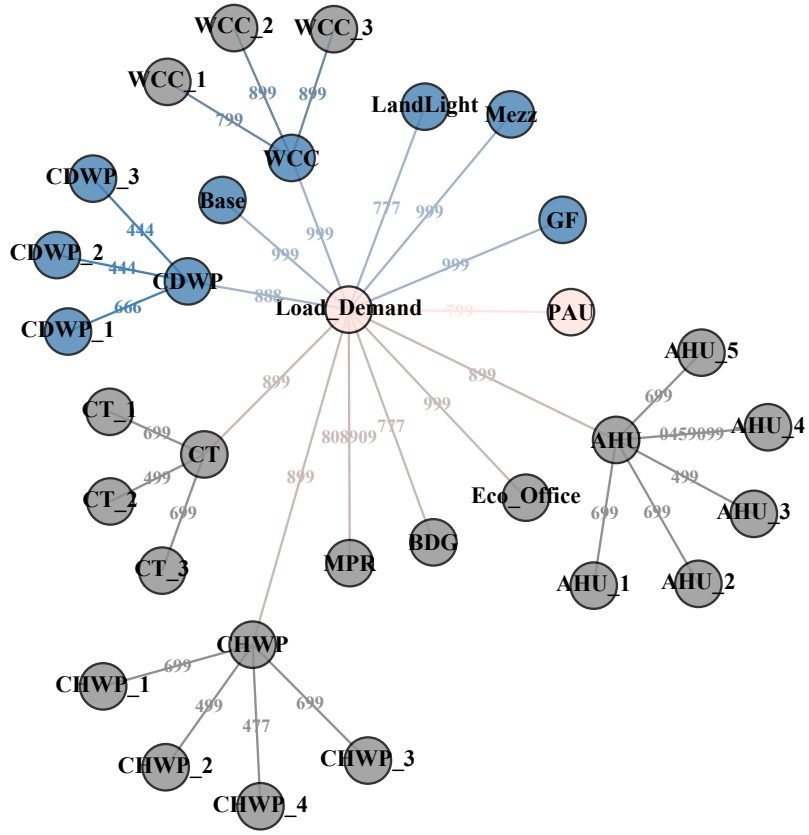


Figure 6.13 An atypical operation on July 2, 2013 (Tuesday)

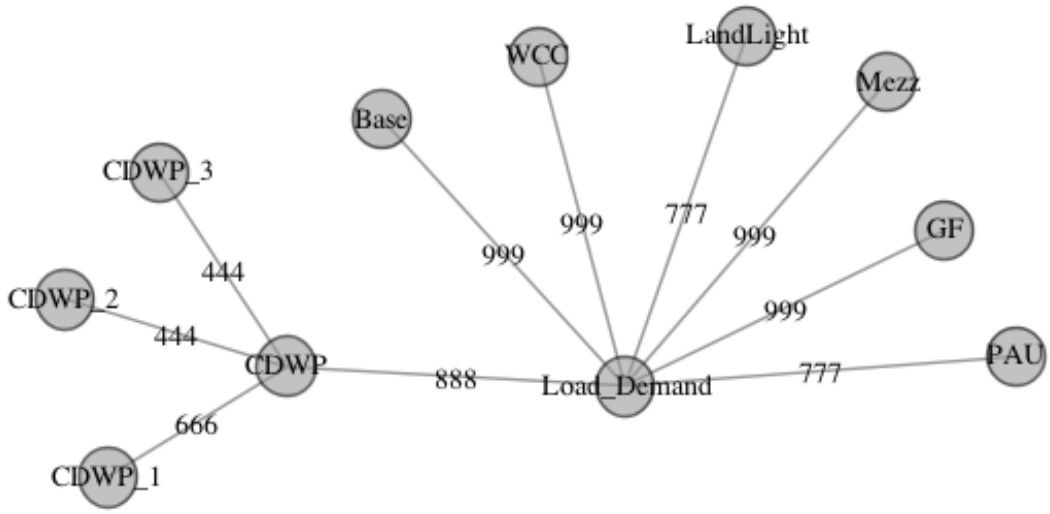


Figure 6.14 The frequent subgraph considered in Figure 6.13

Figure 6.15 presents an atypical operation on September 2, 2013 (Monday) with reference to the frequent subgraph shown in Figure 6.16. It is observed that the power

consumption of landscape lighting was "*Low*" and "*High*" at 17:00 and 18:00, while "Idle" in the frequent subgraph considered. Further inspection shows that the landscape lighting during hot seasons generally operates between 19:00 to 7:00. Such atypical operation can be caused by faults in manual control or poor outdoor visibility.

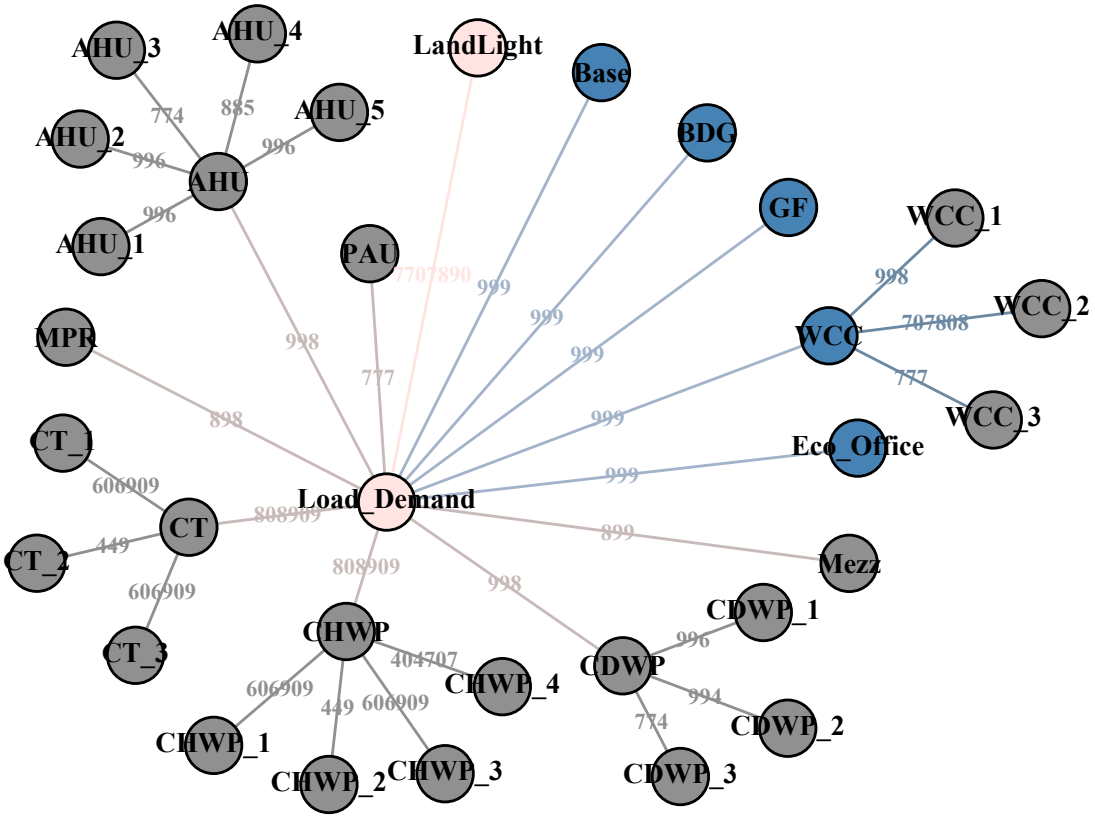


Figure 6.15 An atypical operation on September 2, 2013 (Monday)

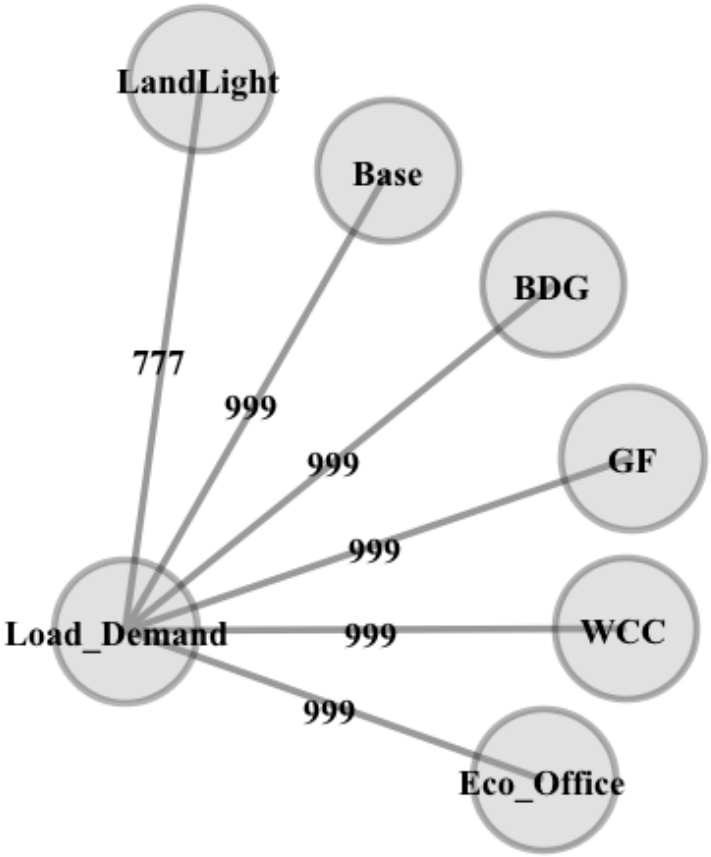


Figure 6.16 The frequent subgraph considered in Figure 6.15

6.4 Summary

With the advances in technologies, building data being collected will not only expand in data volumes, but also have more complex data structures and data types. The effective mining of building data requires a more efficient way for integrating and representing complex information. The majority of DM techniques require the data to be prepared in a rather simple data structure, i.e., a single two-dimensional data table. Meanwhile, typical data preprocessing tasks, such as table joining, cannot fully fulfill the needs of integrating data with complex structures and various types.

This chapter proposes a graph-based data mining methodology as a general solution to ensure the efficiency and effectiveness in knowledge discovery. The reasons are twofold. Firstly, graph provides a flexible way to represent data with potentially complex structures or from multiple sources, e.g., multi-relational data. It is capable of describing complicated relationships by establishing nodes and links, specifying their labels associated and etc. Users could design various graph formats to meet their needs, e.g., the type of knowledge to be discovered. Secondly, the knowledge discovered by graph-based data mining techniques is also represented as graphs. Such knowledge representation has high interpretability and therefore, can greatly reduce the difficulties in the post-mining stage.

The graph-based mining methodology is developed with the intention to discover potentially useful and previous unknown knowledge. The frequent subgraph mining (FSM) is selected as key knowledge discovery technique. Two challenges are specifically addressed. The first is to develop methods to efficiently represent BAS data as graphs. Two graph generation methods, i.e., the observation-based and variable-based methods, are developed. The observation-based method focuses on preserving the low-level detailed information while the variable-based method aims to extract high-level abstracts of BAS data. The second challenge is to ensure the efficiency in post-mining given large amounts of knowledge discovered. Two post-mining methods, i.e., pattern summarization and graph-based anomaly detection, are developed for knowledge selection and application. The pattern summarization

method adopts clustering analysis and graph edit-distance to derive representative patterns. The number of representative patterns is much smaller than the number of frequent subgraphs discovered. It helps users to quickly grasp the essentials of knowledge discovered. The graph-based anomaly detection method uses the frequent subgraphs discovered as a knowledge database, based on which anomaly scores are computed for new observations. The methodology has been applied to mine the BAS data retrieved from a building in Hong Kong. Useful knowledge has been discovered to understand the system operation behaviors and identify atypical operations. The open-source software *R*, *ParSeMis* and *Gephi* were used to perform the mining and visualization tasks.

Advanced data analytics and domain expertise are the both indispensable to ensure the energy efficiency in building operations. This research serves as an exploratory study to investigate the usefulness of graph-based data mining in the knowledge discovery from building data. One limitation is that as the current FSM techniques only work with graphs labeled with categorical values, data transformation is inevitable and the information loss associated is difficult to control. One solution is to develop advanced FSM algorithms which are compatible with numerically weighted graphs. The other is to develop suitable data transformation methods for different types of BAS variables, e.g., temperature, flow-rates and power consumptions for a diversity of building services components.

CHAPTER 7 CONCLUSIONS AND RECOMMENDATAIONS

Building energy efficiency plays a vital role in global energy conservation and environmental sustainability. The development in building technologies has made modern buildings become not only energy intensive, but also information intensive. The knowledge hidden in the vast amount of building data can bring significant benefits in understanding the building operation behaviors, evaluating services system performance, detecting possible faults in operations, and spotting opportunities for energy conservation. Data mining-based big data analysis is a promising approach for knowledge discovery and has gained great success in many industries. It has the ability to discover previous unknown yet potentially useful insights from massive data. Nevertheless, the efficient and effective utilization of data mining techniques in analyzing building data is a non-trivial task. Currently, there is a knowledge gap between building professionals and advanced data mining techniques. Building professionals used to solve building engineering problems based on physical laws and domain experience. Such approach lacks effectiveness when handling massive amounts of data. The success implementation of advanced data mining techniques requires a thoroughly designed analytic process, which accounts for a diversity of challenges, such as the complexity and poor quality in BAS data, the selection of

suitable algorithms for various tasks, the challenges in post-mining massive amounts of knowledge discovered, and the generalization performance on different buildings. Currently, very few research outcomes and practical experiences are available for the effective use of the advanced big data analytics in building management. The research work presented in this thesis has addressed the need through making the following contributions.

Conclusions on Main Contributions

i. This research develops a generic data mining-based analytic framework for the knowledge discovery from the big building operational data as well as applications of the knowledge discovered in building energy management. The framework serves as a prototype of developing methodologies for discovering and applying various types of knowledge from the building operational data. The framework is deliberately designed with considerations of the essential steps in knowledge discovery process, the unique characteristics of building operational data, and the potential tasks in building energy management.

ii. A generic methodology has been developed to discover cross-sectional knowledge in building operational data. Cross-sectional knowledge refers to the relationships and associations between variables without taking into account the temporal dependency. The methodology enables the discovery of previously unknown yet potentially useful insights from massive building operational data.

iii. A generic methodology has been developed to discover temporal knowledge in building operational data. To the best of the author's knowledge, it is the first unsupervised data mining-based methodology developed for exploring temporal knowledge in building operations. The methodology enables the discovery of new types of knowledge, which is represented as sequential motifs and temporal associations. The knowledge discovered can better characterize the building system dynamics and describe the complex interactions between variables in building operations.

iv. A graph-based data mining methodology is developed to tackle the new challenges brought by the advance in building technologies, i.e., data are being collected throughout the whole building lifecycle, and are of different types (e.g., text data, video data and numeric data) and structures (e.g., multi-relational databases). Conventional methodologies, which are typically designed for analyzing data stored in a single two-dimensional data table, are of limited use when building data become more complex in data types and structures. This study has developed general solutions for the effective transformation from building data into graphs, knowledge discovery based on graph data, and post-mining of the graph-based knowledge discovered. The usefulness of this methodology has been validated through case studies.

Summary of The Data Mining-based Analytic Framework

Based on a comprehensive exploration of advanced data mining techniques, in-depth analysis of building operational data characteristics as well as considerations for practical applications, a generic data mining-based analytic framework is developed. The framework consists of 4 major phases, i.e., data exploration, data partitioning, knowledge discovery and post-mining.

The data exploration contains two tasks, i.e., data visualization and data preprocessing. Three subtasks are included in the data preprocessing step, including the data cleaning, data transformation and data reduction. The data partitioning phase aims to discover the intrinsic characteristics in building operational data according to the building energy consumptions. Building operational data are usually highly dynamic and the values of a variable may vary greatly under different operation conditions. The inclusion of the data partitioning phase helps to enhance the reliability and quality of the knowledge discovered in the following phases. The knowledge discovery phase adopts a diversity of advanced data mining techniques to mine potentially useful knowledge. Considering that building data are typically stored in a single two-dimensional data table, two types of knowledge can be discovered based on the axis to explore, i.e., cross-sectional knowledge (e.g., associations between

variables at the same time step) and temporal knowledge (e.g., associations between variables at different time steps). This research also develops a third methodology considering that future building data may have complex data structures, e.g., multi-relational databases. Given the large amounts of knowledge discovered, the fourth phase, i.e., post-mining, is designed to improve the efficiency in knowledge selection, transformation and interpretation.

Summary of The Methodology for Cross-sectional Knowledge Discovery

A methodology for mining cross-sectional knowledge is developed based on the generic DM-based analytic framework. The association rule mining (ARM) and the quantitative association rule mining (QARM) are adopted as the main techniques used at the knowledge discovery phase. The methods adopted and developed at the other phases are deliberately designed considering the building operational data characteristics and the compatibility with ARM and QARM. Two indices of high practical values are defined to facilitate the post-mining of QARM, i.e. the standard deviation of lift (SD-Lift) of rules with similar rule pattern and the abnormality degree (AD). SD-Lift can help to fast select useful rules from a large number of rules obtained in QARM, which is a major obstacle to the application of QARM. AD provides a generic method of using the association rules for detecting abnormalities.

The methodology has been validated through the use of real-world data retrieved from the International Commerce Center (ICC) in Hong Kong. The results obtained

are very encouraging. The change of operation strategy, non-typical and abnormal operations and sensor fault occurring during operation in ICC air conditioning system are successfully detected and diagnosed.

Summary of The Methodology for Temporal Knowledge Discovery

A methodology for mining temporal knowledge is developed based on the generic DM-based analytic framework. Motif discovery and temporal association rule mining (QARM) are adopted as the main techniques used for knowledge discovery. The methodology specifically addresses two major challenges in mining temporal knowledge. One is the heavy computational load caused by the massive data amount. Period estimation and data transformation are integrated into the data exploration phase to tackle this challenge. The other is the efficient utilization of knowledge discovered. Two methods are developed at the post-mining stage. The first uses a co-occurrence matrix to map the relationship between univariate motifs. Reliable associations between univariate motifs are derived which provides a novel and convenient approach to utilizing univariate motifs. The second method utilizes a filtering method to improve the temporal association rules mining algorithms with the accurate estimation of time interval between the antecedent and the consequent. The time interval or lag provides valuable insights into building dynamics and HVAC performance characteristics. The effectiveness of the methodology has been validated

using the data retrieved from ICC. The knowledge discovered has been successfully used to identify anomalies in building operations and characterize the building dynamics.

Summary of The Methodology for Mining BAS Data with Complex Data Structures

A graph-based data mining methodology is developed for the knowledge discovery from building operational data with complex data structures, e.g., multi-relational databases. It aims to provide a general solution to the knowledge discovery from future building data, which may be stored not only in a single two-dimensional data table, but also in other formats, e.g., building information models. In view of this, the graph-based data mining technology is also developed based on the generic framework. The frequent subgraph mining (FSM) is selected as the main mining technique. Two graph generation methods, i.e., the observation-based and variable-based methods, are developed to transform building data into graphs. Two post-mining methods, i.e., pattern summarization and graph-based anomaly detection, are developed for efficiency in the post-mining stage. The methodology has been applied to mine the building operational data retrieved from the Zero Carbon Building (ZCB) in Hong Kong. Valuable knowledge has been discovered to understand the system operation behaviors and identify atypical operations.

Recommendations for Future Work

Major efforts of this thesis are made on the development of the data mining-based analytic framework and methodologies for the knowledge discovery from building operational data. It is desirable and valuable to make further efforts on the following three aspects related to the research presented in this thesis.

- This research has developed a graph-based data mining methodology for mining building data with complex data structures. Due to the data availability, the methodology has been validated using a building operational data set which mainly consists of the measurement data. The potential applications of this methodology can be more thoroughly presented if it is integrated with building information models. Further study will be performed to investigate the usefulness of the methodology in analyzing the complex information stored in building information models.
- The software packages for the framework proposed in this research need to be developed for the integration with building automation systems. The packages should be designed considering the convenience of building operators. It should provide a user-friendly interface which enables building operators to easily perform different types of mining tasks. In addition, it should also provide a development environment for building operators to conveniently write add-ons for possible extensions.
- This research focuses on the knowledge discovery from structured data retrieved

from building operations. There is a large amount of unstructured data which can be obtained from daily building operations, such as video data and text data. The video data refer to the information collected by the closed circuit television (CCTV). Such data can be applied to derive various types of knowledge, especially on occupant behaviors and indoor environment. The text data is another main information source existed in building operations. Text mining techniques can be applied extract useful knowledge from texts. It may provide additional useful knowledge for building energy management. So far, the potential of knowledge discovery from unstructured data in building operations has not been explored. It can be an interesting and promising direction for future study.

REFERENCES

Acar, U.A., Ihler, A., Mettu, R., Sumer, O., Adaptive Bayesian inference, Neural Information Systems (NIPS), 2007.

Agrawal, R., Srikant, R., Fast algorithms for mining association rules, In: Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.

Ahmad, A.S., Hassan, M.Y., Abdullah, M.P., Rahman, H.A., Hussin, F., Abdullah, H., Saidur, R., A review on applications of ANN and SVM for building electrical energy consumption forecasting, Renewable and Sustainable Energy Reviews 33 (2014) 102-109.

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

Ahmed, A., Ploennigs, J., Menzel, K., Cahill, B., Multi-dimensional building performance data management for continuous commissioning, Advanced Engineering Informatics 34 (4) (2010) 466-475.

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

Apte, C., Weiss, S., Data mining with decision trees and decision rules, Future Generation Computer Systems 13 (2-3) (1997) 197-210.

Ardakani, A.J., Ardakani, F.F., Hosseinian, S.H., A novel approach for chiller loading using particle swarm optimization, Energy and Buildings 40 (12) (2008) 2177-2187.

Azadeh, A., Saberi, M., Ghaderi, S.F., Gitiforouz, A., Ebrahimipour, V., Improved estimation of electricity demand function by integration of fuzzy system and data mining approach, Energy Conversion and Management 49 (8) (2008) 2165-2177.

Bailey, M.B., Kreider, J.F., Creating an automated chiller fault detection and

diagnostics tool using a data fault library, ISA Transactions 42 (3) (2003) 485-495.

Beghi, A., Cecchinato, L., Cosi, G., Rampazzo, M., A PSO-based algorithm for optimal multiple chiller systems operation, Applied Thermal Engineering 32 (2012) 31-40.

Beghi, A., Cecchinato, L., Rampazzo, M., A multi-phase genetic algorithm for the efficient management of multi-chiller systems, Energy Conversion and Management 52 (3) (2011) 1650-1661.

Bendapudi, S., Braun, J.E., A review of literature on dynamic models for vapor compression equipment, 2002, Report 4036-5, Ray Herrick Laboratories, Purdue University.

Benezeth, Y., Laurent, H., Emile, B., Rosenberger, C., Towards a sensor for detecting human presence and characterizing activity, Energy and Buildings 43 (2-3) (2011) 305-314.

Ben-Nakhi, A.E., Mahmoud, M.A., Cooling load prediction for buildings using general regression neural networks, Energy Conversion and Management 45 (13-14) (2004) 2127-2141.

Bichiou, Y., Krarti, M., Optimization of envelope and HVAC systems selection for residential buildings, Energy and Buildings 43 (12) (2011) 3373-3382.

Borgelt C., Berthold M. Mining molecular fragments: Finding relevant substructures of molecules. Proceedings of International Conference on Data Mining, 2002, 211-218.

Bornatico, R., Pfeiffer, M., Witzig, A., Guzzella, L., Optimal sizing of a solar thermal building installation using particle swarm optimization, Energy 41 (1) (2012) 31-37.

Breiman, L., Random forests, Machine Learning 45 (1) (2001) 5-32.

Cabrera, D.F.M., Zareipour, H., Data association mining for identifying lighting energy waste patterns in educational institutes, Energy and Buildings 62 (2013)

210-216.

Cao, K., Huang, B., Wang, S., Lin, H., Sustainable land use optimization using boundary-based fast genetic algorithm, Computers, Environment and Urban System 36 (3) (2012) 257-269.

Capozzoli, A., Lauro, F., Khan, I., Fault detection analysis using data mining techniques for a cluster of smart office buildings, Expert Systems with Applications 42 (9) (2015) 4324-4338.

Cassol, F., Schneider, P.S., Franca, F.H.R., Silva Neto, A.J., Multi-objective optimization as a new approach to illumination design of interior spaces, Building and Environment 46 (2) (2011) 331-338.

Castro, N., Performance evaluation of a reciprocating chiller using experimental data and model predictions for fault detection and diagnosis, ASHRAE Transactions 108 (1) (2002) 889-903.

Cerovsek, T., A review and outlook for a "Building Information Model" (BIM): A multi-standpoint framework for technological development, Advanced Engineering Informatics 25 (2) (2011) 224-244.

Chang, Y.C., Genetic algorithm based optimal chiller loading for energy conservation, Applied Thermal Engineering 25 (17-18) (2005) 2800-2815.

Chang, Y.C., Lee, C.Y., Chen, C.R., Chou, C.J., Chen, W.H., Chen, W.H., Evolution strategy based optimal chiller loading for saving energy, Energy Conversion and Management 50 (1) (2009) 132-139.

Chang, Y.C., Lin, J.K., Chuang, M.H., Optimal chiller loading by genetic algorithm for reducing energy consumption, Energy and Buildings (37) (2) (2005) 147-155.

Chen, C.L., Chang, Y.C., Chan, T.S., Applying smart models for energy saving in optimal chiller loading, Energy and Buildings 68 (A) (2014) 364-371.

Chen, Y.M., Hao, X.L., Zhang, G.Q., Flow meter fault isolation in building central chilling systems using wavelet analysis, Energy Conversion and Management 47

(13-14) (2006) 1700-1710.

Cheng, C.W., Leu, S.S., Cheng, Y.M., Wu, T.C., Lin, C.C., Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry, Accident Analysis & Prevention 48 (12) (2012) 214-222.

Cheng, C.W., Lin, C.C., Leu, S.S., Use of association rules to explore cause-effect relationships in occupational accidents in the Taiwan construction industry, Safety Science 48 (4) (2010) 436-444.

Cheng, M.Y., Hoang, N.D., Wu, Y.W., Hybrid intelligence approach based on LS-SVM and differential evolution for construction cost index estimation: A Tawiwan case study, Automation in Construction 35 (35) (2013) 306-313.

Chi, S., Suk, S.J., Kang, Y., Mulva, S.P., Development of a data mining-based analysis framework for multi-attribute construction project information, Advanced Engineering Informatics 26 (3) (2012) 574-581.

Chiu B., Keogh E., Lonardi S., Probabilistic discovery of time series motifs, ACM SIGKDD, Washington, DC, USA, 2003, pp. 493-498.

Choiniere D., Corsi, M., A BEMS-assisted commissioning tool to improve the energy performance of HVAC systems, In Proceedings of the ICEBO 2003, Berkeley, California; October 13–15, 2003.

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

Chow, T.T., Zhang, G.Q., Lin, Z., Song, C.L., Energy and Buildings 34 (1) (2002) 103-109.

Clarke, J., McLay, L., McLeskey Jr., J.T., Comparison of genetic algorithm to particle swarm for constrained simulation-based optimization of a geothermal power plant, Advanced Engineering Informatics 28 (1) (2014) 81-90.

Construction Industry Council (CIC), ZCB fact sheet, 2002.

Cook D.J., Holder L.B. Graph-based data mining. IEEE Intelligent Systems 15 (2)

(2000) 32-41.

Cook D.J., Holder L.B. Substructure discovery using minimum description length and background knowledge. Journal of Artificial Intelligence Research 1 (1994) 231-255.

Cook, D.J., Holder, L.B., Mining graph data. 1st edition. 2006. Wiley-Interscience, New Jersey, USA.

Cortes, C., Vapnik, V., Support vector networks, Machine Learning 20 (3) (1995) 273-297.

Dalene, F., Technology and information management for low-carbon building, Journal of Renewable Sustainable Energy 4 (2012) 041402; doi: 10.1063/1.3694120.

Daw C.S., Finney C.E.A., Tracy E.R., A review of symbolic analysis of experimental data, Review of Scientific Instruments 74 (2003) 915-930.

DeRosa, M., Data mining and data analytics for counterterrorism, Center for Strategic & International Studies, Washington DC, USA, 2004.

Dietterich, T.G., Ensemble methods in machine learning, In: Multiple Classifier Systems, vol.1857, 1-15, Springer, 2001.

Djuric, N., Novakovic, V., Review of possibilities and necessities for building lifetime commissioning, Renewable and Sustainable Energy Reviews 13 (2009) 486-492.

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

Dos Santos Ceolho, L., Mariani, V.C., Improved firefly algorithm approach applied to chiller loading for energy conversion, Energy and Buildings 59 (2013) 273-278.

Du, Z.M., Fan, B., Jin, X.Q., Chi, J.L., Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis, Building and Environment 73 (2014) 1-11.

Du, Z.M., Jin, X.Q., Multiple faults diagnosis for sensors in air handling unit using Fisher discriminant analysis, Energy Conversion and Management 49 (12) (2008) 3654-3665.

Du, Z.M., Jin, X.Q., Wu, L.Z., PCA-FDA-based fault diagnosis for sensors in VAV systems, HVAC&R Research 13 (2) (2007) 349-367.

Electrical & Mechanical Services Department (EMSD), Hong Kong Energy End-use Data 2012, The Government of the Hong Kong Special Administrative Region, 2012.

Erickson, V.L., Lin, Y.Q., Kamthe, A., Brahme, R., Surana, A., Cerpa, A.E., Sohn, M.D., Narayanan, S., Energy efficient building environment control strategies using real-time occupancy measurements, In Proceedings of the 1st ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, 2009, 19-24.

Evins, R., A review of computational optimization methods applied to sustainable building design, Renewable and Sustainable Energy Reviews 22 (11) (2013) 230-245.

Fan C., Xiao F., Yan C.C., A framework for knowledge discovery in massive building automation data and its applications in building diagnostics, Automation in Construction 50 (2015) 81-90.

Fan, C., Xiao, F., Wang, S.W., Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques, Applied Energy 127 (C) (2014a) 1-10.

Fan, C., Xiao, F., Wang, S.W., Rare event analysis of high dimensional building operational data using data mining techniques, The 3rd International High Performance Buildings Conference at Purdue, July 14-17, 2014b, Indiana, USA.

Fan, C., Xiao, F., Yan, C.C., A framework for knowledge discovery in massive building automation data and its application in building diagnostics, Automation in Construction 50 (8) (2015) 81-90.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., The KDD process for extracting useful knowledge from volumes of data, Communication of ACM 39 (11) (1996) 27-34.

Fred A.L.N., Jain A.K., Combining multiple clustering using evidence accumulation, IEEE Transactions on Pattern Analysis and machine intelligence 27 (2005) 835-850.

Fu T.C., A review on time series data mining, Engineering Applications of Artificial Intelligence 17 (2011) 164-181.

Gagne, J.M.L., Andersen, M., Multi-objective façade optimization for daylighting design using a genetic algorithm, 2010, The 4th National Conference on IBPSA, New York, USA.

Gantz, J., Reinsel, D., The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, International Data Corporation, IDC iView: IDC Analyze the Future, 2012.

Gao D.C., Wang S.W., Sun Y.J., A fault-tolerant and energy efficient control strategy for primary-secondary chilled water systems in buildings, Energy and Buildings 43 (2011) 3646-3656.

Gao X.B., Xiao B., Tao D.C., Li X.L. A survey of graph edit distance. Pattern Analysis and Applications 13 (2010) 113-129.

Gershenson, C., Artificial neural networks for beginners, Cognitive and Computing Sciences, 2003, University of Sussex, UK.

Ginestet, S., Marchio, D., Morisot, O., Improvement of buildings energy efficiency: Comparison, operability and results of commissioning tools, Energy Conversion and Management 76 (2013) 368-376.

Ginestet, S., Marchio, D., Retro and on-going commissioning tool applied to an existing building: Operability and results of IPMVP, Energy 35 (4) (2010) 1717-1723.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.,

Detecting influenze epidemics using search engine query data, Nature 457 (2009) 1012-1014.

Goia, F., Haase, M., Perino, M., Optimizing the configuration of a façade module for office buildings by means of integrated thermal and lighting simulations in a total energy perspective, Applied Energy 108 (C) (2013) 515-527.

Grira, N., Crucianu, M., Boujemaa, N., Unsupervised and semi-supervised clustering: A brief survey, The 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, 2005, 9-16, New York, USA.

Gulbinas R., Khosrowpour A., Taylor J., Segmentation and classification of commercial building occupants by energy-use efficiency and predictability, IEEE Transactions on Smart Grid 6 (2015) 1414-1424.

Gupta M., Gao J., Aggarwal C.C., Han J.W., Outlier detection for temporal data: A survey, IEEE Transactions on Knowledge and Data Engineering 15 (2014) 1-20.

Guyon, I., Elisseeff, A., An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157-1182.

Han, J.W., Kamber, M., Data mining: Concepts and techniques, 3rd edition, The Morgan Kaufmann Series in Data Management Systems, 2011.

Han, J.W., Pei, J., Yin, Y.W., Mining frequent patterns without candidate generation, Technical Report CMPT99-12, School of Computing Science, Simon Fraser University, 1999.

Hang, Y., Du, L.L., Qu, M., Peeta, S., Multi-objective optimization of integrated solar absorption cooling and heating systems for medium-sized office buildings, Renewable Energy 52 (6) (2013) 67-78.

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

He, X.F., Zhang, Z.J., Kusiak, A., Performance optimization of HVAC systems with computational intelligence algorithms, Energy and Buildings 81 (2014) 371-380.

Holst, J.N., Using whole building simulation models and optimizating procedures to optimize building envelope design with respect to energy consumption and indoor environment, In Proceedings of the 8th International IBPSA Conference, 2003, Eindhoven, Netherlands.

Hong Kong energy end-use data 2012, Electrical & Mechanical Services Department, HKSARS, September 2012.

Hothorn, T., Hornik K., Zeileis A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15 (3) (2006) 651-674.

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

Hou, Z.J., Lian, Z.W., Yao, Y., Yuan, X.J., Cooling-load prediction by the combination of rough set theory and an artificial neural-network based on data-fusion techniques, Applied Energy 83 (9) (2006) 1033-1046.

House, J.M., Vaezi-Najad, H., Whitcomb, J.M., An expert rule set for fault detection in air-handling units, ASHRAE Transactions 107 (1) (2001) 858-871.

House, J.M., Vaezi-Nejad, H., Whitcomb, J.M., An expert rules set for fault detection in air-handling units/discussion, ASHRAE Transactions 107 (1) (2001) 858-871.

Hu, Y.P., Chen, H.X., Xie, J.L., Yang, X.S., Zhou, C., Chiller sensor fault detection using a self-adaptive principal component analysis method, Energy and Buildings 54 (2012) 252-258.

Huan J., Wang W., Prins J. Efficient mining of frequent subgraph in the presence of isomorphism. Proceedings of the 2003 International Conference on Data Mining, 2003, 549-552.

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

Inokuchi A., Washio T., Motoda H. An Apriori-based algorithm for mining frequent substructures from graph data. Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2000, 13-23.

International Energy Agency (IEA), 2015, accessed on October 28, https://www.iea.org/aboutus/faqs/energyefficiency/

Jain, A., Satish, B., Clustering based short term load forecasting using support vector machines, In Proceedings of the IEEE Power Tech Conference, June 28 – July 2, 2009, Bucharest, Romania.

Jakkula, V., Cook, D., Outlier detection in smart environment structured power datasets, In Proceedings of the IEEE Intelligent Systems, London, UK, 2010.

Janetzko, H., Stoffel, F., Mittelstadt, S., Keim, D.A., Anomaly detection for visual analytics of power consumption data, Computer & Graphics 38 (2014) 27-37.

Jetcheva, J.G., Majidpour, M., Chen, W.P., Neural network model ensembles for building-level electricity load forecasts, Energy and Buildings 84 (2014) 214-223.

Jia, Y., Reddy, T.A., Characteristics physical parameter approach to modeling chillers suitable for fault detection, diagnosis and evaluation, ASME Journal of Solar Energy Engineering 125 (3) (2003) 258-265.

Jiang C.T., Coenen F., Zito M. A survey of frequent subgraph mining algorithms. The Knowledge Engineering Review (0) (2004) 1-31.

Jin, X.Q., Du, Z.M., Fault tolerant control of outdoor air and AHU supply air temperature in VAV air conditioning systems using PCA method, Applied Thermal Engineering 26 (11-12) (2006) 1226-1237.

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

Jota, P.R.S., Silva, V.R.B., Jota, F.G., Building load management using cluster and statistical analyses, International Journal of Electrical Power & Energy Systems

33 (8) (2011) 1498-1505.

Karatasou, S., Santamouris, M., Geros, V., Modeling and predicting building energy use with artificial neural networks: Methods and results, Energy and Buildings 38 (8) (2006) 949-958.

Katipamula, S., Brambley, M.R., Methods for fault detection, diagnostics, and prognostics for building systems-a review, Part I, HVAC&R Research 11 (1) (2005)(a) 3-25.

Katipamula, S., Brambley, M.R., Methods for fault detection, diagnostics, and prognostics for building systems-a review, Part II, HVAC&R Research 11 (2) (2005)(b) 169-187.

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

Kim, G., Schaefer, L., Lim, T.S., Kim, J.T., Thermal comfort prediction of an underfloor air distribution system in a large indoor environment, Energy and Buildings 64 (2013) 323-331.

Kim, H., Anderson, K., Lee, S.H., Hildreth, J., Generating construction schedules through automatic data extraction using open BIM (building information modeling) technology, Automation in Construction 35 (2013) 285-295.

Kim, H., Stumpf, A., Kim, W., Analysis of an energy efficient building design through data mining approach, Automation in Construction 20 (1) (2011) 37-43.

Koegh, E., Lin, J., Lee, S.H., Van Herle, H., Finding the most unusual time series subsequence: algorithms and applications, Knowledge and Information Systems 11 (1) (2006) 1-27.

Koh, H.C., Tan, G., Data mining applications in healthcare, Jounral of Healthcare Information Management 19 (2) (2005) 64-72.

Kolter, J.Z., Ferreira, J., A large-scale study on predicting and contextualizing building energy use, In Proceedings of the 25th AAAI Conference on Artificial

Intelligence, Aug 7-11, 2011, San Francisco, California, USA.

Kruger, E., Givoni, B., Thermal monitoring and indoor temperature predictions in a passive solar building in an arid environment, Building and Environment 43 (11) (2008) 1792-1804.

Kucuksille, E.U., Selbas, R., Sencan, A., Data mining techniques for properties of refrigerants, Energy Conversion and Management 50 (2) (2009) 399-412.

Kucuksille, E.U., Selbas, R., Sencan, A., Prediction of thermodynamic properties of refrigerants using data mining, Energy Conversion and Management 52 (2) (2011) 836-848.

Kumar, R., Sinha, A.K., Singh, B.K., Modhukalya, U., A design optimization tool of earth-to-air heat exchanger using a genetic algorithm, Renewable Energy 33 (10) (2008) 2282-2288.

Kuramochi M., Karypis G. GREW-A scalable frequent subgraph discovery algorithm. Proceedings of the 4th IEEE International Conference on Data Mining, 2004, 439-442.

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

Kusiak, A., Li, M.Y., Zhang, Z.J., A data-driven approach for steam load prediction in buildings, Applied Energy 87 (3) (2010) 925-933.

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

Kusiak, A., Xu, G.L., Tang, F., Optimization of an HVAC system with a strength multi-objective particle-swarm algorithm, Energy 36 (10) (2011) 5935-5943.

Kwac J., Flora J., Rajagopal R., Household energy consumption segmentation using hourly data, IEEE Transactions on Smart Grid 5 (2014) 420-430.

Kwok, S.K., Yuen, K.K., Lee, W.M., An intelligent approach to assessing the effect of building occupancy on building cooling load prediction, Building and Environment 46 (8) (2011) 1681-1690.

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

Lazarevic, A., Srivastava J., Kumar V. Data mining for analysis of rare events: A case study in security, financial and medical applications. The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) Tutorial, 2004.

Lee, T.S., Lu, W.C., An evaluation of empirically-based models for predicting energy performance of vapor-compression water chillers, Applied Energy 87 (11) (2010) 3486-3493.

Lee, W.S., Chen, Y.T., Kao, Y.C., Optimal chiller loading by differential evolution algorithm for reducing energy consumption, Energy and Buildings 43 (2-3) (2011) 599-604.

Lee, W.S., Lin, L.C., Optimal chiller loading by particle swarm algorithm for reducing energy consumption, Applied Thermal Engineering 29 (8-9) (2009) 1730-1734.

Lee, W.Y., House, J.M., Kyong, N.H., Subsystem level fault diagnosis of a building's air-handling unit using general regression neural networks, Applied Energy 77 (2) (2004) 153-170.

Leskovar, V.Z., Premrov, M., An approach in architectural design of energy-efficient timber buildings with a focus on the optimal glazing size in the south-oriented façade, Energy and Buildings 43 (12) (2011) 3410-3418.

Li, Q., Meng, Q.L., Cai, J.J., Yoshino, H., Mochida, A., Applying support vector machine to predict hourly cooling load in the building, Applied Energy 86 (10) (2009) 2249-2256.

Liang, J., Du, R., Model-based fault detection and diagnosis of HVAC systems using support vector machine method, International Journal of Refrigeration 30 (6) (2007) 1104-1114.

Liao, C.W., Perng, Y.H., Data mining for occupational injuries in the Taiwan

construction industry, Safety Science 46 (7) (2008) 1091-1102.

Lin, J., Keogh, E., Wei, L., Lonardi, S., Experiencing SAX: A novel symbolic representation of time series, Data Mining and Knowledge Discovery 15 (2) (2007) 107-144.

Lu, L., Cai, W.J., Xie, L.H., Li, S.J., Soh, Y.C., HVAC system optimization: In-building section, Energy and Buildings 37 (2) (2005) 11-22.

Lu. L., Cai, W.J., Chai, Y.S., Xie, L.H., Global optimization for overall HVAC systems-Part I problem formulation and analysis, Energy Conversion and Management 46 (7) (2005a) 999-1014.

Lu. L., Cai, W.J., Chai, Y.S., Xie, L.H., Global optimization for overall HVAC systems-Part II problem solution and simulations, Energy Conversion and Management 46 (8) (2005b) 1015-1028.

Ma, Z.J., Wang, S.W., Building energy research in Hong Kong: A review, Renewable and Sustainable Energy Reviews 13 (8) (2009) 1870-1883.

Machairas, V., Tsangrassoulis, A., Axarli, K., Algorithms for optimization of building design: A review, Renewable and Sustainable Energy Reviews 31 (C) (2014) 101-112.

Madsen H., Time series analysis, 1st edition, Chapman & Hall/CRC Texts in Statistical Science, 2007.

Magnier, L., Haghighat, F., Multiobjective optimization of building design using TRNSYS simulation, genetic algorithm, and artificial neural network, Building and Environment 45 (3) (2010) 739-746.

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

Maimon O., Rokach L., Data mining and knowledge discovery handbook, 2nd edition, Springer, New York, 2010.

Mena, R., Rodriguez, F., Castilla, M., Arahal, M.R., A prediction model based on

neural networks for the energy consumption of a bioclimatic building, Energy and Buildings 82 (2014) 142-155.

Meta, J., Alvarez, J.L., Riquelme, J.C., An evolutionary algorithm to discover numeric association rules, In Proceedings of the ACM Symposium on Applied Computing, 2002, 590-594.

Mikut, R., Reischl, M., Data mining tools, Data Mining and Knowledge Discovery 1 (5) (2011) 431-443.

Miller C., Nagy Z., Schlueter A., Automated daily pattern filtering of measured building performance data, Automation in Construction 49 (2015) 1-17.

Minnen K., Isbell C.L., Essa I., Starner T., Discovering multivariate motifs using subsequence density estimation and greedy mixture learning, The 22nd National Conference on Artificial Intelligence, Volumn 1, 2007, pp. 615-620.

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

Mustafaraj, G., Chen, J., Lowry, G., Thermal behavior prediction utilizing artificial neural networks for an open office, Applied Mathematical Modeling 34 (11) (2010) 3216-3230.

Nagi, J., Yap, K.S., Nagi, F., Tiong, S.K., Ahmed, S.K., A computational intelligence scheme for the prediction of daily peak load, Applied Soft Computing 11 (8) (2011) 4773-4788.

Ngai, E.W.T., Xiu, L., Chau, D.C.K., Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications 36 (2) (2009) 2592-2602.

Nijssen S., Kok J.N. A quickstart in frequent structure mining can make a difference. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, 647-652.

Norford, L.K., Wright, J.A., Buswell, R.A., Luo, D., Klaassen, C.J., Suby, A., Demonstration of fault detection and diagnosis methods for air-handling units

(ASHRAE 1020-RP), HVAC&R Research 8 (1) (2002) 41-71.

Ogunsola, O.T., Song, L., Wang, G., Development and validation of a time-series model for real-time thermal load estimation, Energy and Buildings 76 (2014) 440-449.

Olson, D.L., Delen, D., Advanced data mining techniques, Springer, 2008.

Ozcan, H., Ozdemir, K., Ciloglu, Optimum cost of an air cooling system by using differential evolution and particle swarm algorithms, Energy and Buildings 65 (2013) 93-100.

Panapakidis, L.P., Papadopoulos, T.A., Christoforidis, G.C., Papagiannis, G.K., Pattern recognition algorithms for electricity load curve analysis of buildings, Energy and Buildings 73 (2014) 137-145.

Patnaik D., Marwah M., Sharma R.K., Ramakrishnan N., Temporal data mining approaches for sustainable chiller management in data centers, ACM Transactions on Intelligent Systems and Technology 2 (2011) 1-29.

Pearson R.K., Outliers in process modeling and identification, IEEE Transactions on Control Systems Technology 10 (2002) 55-63.

Philippsen M., Worlein M., Dreweke A., Werth T., The parallel and sequential graph mining (ParSeMis) suite, Department Informatik, Friedrich-Alexander Universitat Erlangen-Nurnberg (FAU), German, 2006-2011, https://www2.informatik.uni-erlangen.de/EN/research/zold/ParSeMiS/index.html

Radovanovic, M., Ivanovic, M., Text mining: Approaches and applications, Novi Sad Journal of Mathematics 38 (3) (2008) 227-234.

Ramesh, T., Prakash, R., Shukla, K.K., Life cycle energy analysis of buildings: an overview, Energy and Buildings 42 (10) (2010) 1592-1600.

Reddy, T.A., Andersen, K.K., Pericolo, P.P., Cabrera, G., Evaluation of the suitability of different chiller performance models for on-line training applied to automated fault detection and diagnosis, International Journal of Heating, Ventilating, Air Conditioning and Refrigerating Research 9 (4) (2003) 385-414.

Retkowski, W., Thoming, J., Thermoeconomic optimization of vertical ground-source heat pump systems through nonlinear integer programming, Applied Energy 114 (C) (2014) 492-503.

Rish, I., An empirical study of the naïve bayes classifier, In Proceedings of IJCAI-01 Workshop on Empirical Methods in AI, 41-46, Sicily, Italy, 2001.

Ruano, A.E., Crispim, E.M., Conceicao, E.Z.E., Lucio, M.M.J.R., Prediction of building's temperature using neural networks models, Energy and Buildings 38 (6) (2006) 682-694.

Saeys, Y., Inza, I., Larranaga, P., A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507-2517.

Saez, R.M., Sidrach-de-Cardona, M., Mora-Lopez, L., Data mining and statistical techniques for characterizing the performance of thin-film photovoltaic modules, Expert Systems with Applications 40 (17) (2013) 7141-7150.

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

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

Samatova N.F., Hendrix W., Jenkins J., Padmanabhan K., Chakraborty A. Practical graph mining with R. 1st edition. 2013. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC.

Sayyaadi, H., Amlashi, E.H., Amidpour, M., Multi-objective optimization of a vertical ground source heat pump using evolutionary algorithm, Energy Conversion and Management 50 (8) (2009) 2035-2046.

Schein, J., Bushby, S.T., Castro, N.S., House, J.M., A rule-based fault detection method for air handling units, Energy and Buildings 38 (12) (2006) 1485-1492.

Schlueter A., Thesseling F., Building information model based energy/exergy

performance assessment in early design stages, Automation in Construction 18 (2) (2009) 153-163.

Seem, J.E., Using intelligent data analysis to detect abnormal energy consumption in buildings, Energy and Buildings 39 (1) (2007) 52-58.

Seo, J., Ooka, R., Kim, J.T., Nam, Y., Optimization of the HVAC system design to minimize primary energy demand, Energy and Buildings 76 (2014) 102-108.

Shih, H.C., A robust occupancy detection and tracking algorithm for the automatic monitoring and commissioning of a building, Energy and Buildings 77 (2014) 270-280.

Sivic, J., Zisserman, A., Video data mining using configurations of viewpoint invariant regions, In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, Washington, D.C., USA.

Soibelman, L., Wu, J.F., Caldas, C., Brilakis, I., Lin, K.Y., Management and analysis of unstructured construction data ypes, Advanced Engineering informatics 22 (1) (2008) 15-27.

Son, H., Kim, C., Early prediction of the performance of green building projects using pre-project planning variables: data mining approaches, Journal of Cleaner Production, In Press, doi: 10.1016/j.jclepro.2014.08.071.

Son, H., Kim, C., Kim, C., Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables, Automation in Construction 27 (7) (2012) 60-66.

Stanescu, M., Kajl, S., Lamarche, L., Evolutionary algorithm with three different permutation options used for preliminary HVAC system design, In Proceedings of the 1st Building Simulation and Optimization Conference, 2012, Loughborough, UK.

Strohmeier, S., Piazza, F., Domain driven data mining in human resource management: A review of current research, Expert Systems with Applications 40

(7) (2013) 2410-2420.

Sun, Y.J., Wang, S.W., Huang, G.S., Chiller sequencing control with enhanced robustness for energy efficient operation, Energy and Buildings 41 (11) (2009) 1246-1255.

Talebizadeh, P., Mehrabian, M.A., Abdolzadeh, M., Prediction of the optimum slope and surface azimuth angles using the genetic algorithm, Energy and Buildings 43 (11) (2011) 2998-3005.

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

Tanaka Y., Iwamoto K., Uehara K., Discovery of time-series motif from multi-dimensional data based on MDL principle, Machine Learning 58 (2005) 269-300.

Tang, F., Kusiak, A., Wei, X.P., Modeling and short-term prediction of HVAC system with a clustering algorithm, Energy and Buildings 82 (1) (2014) 310-321.

Tashtoush, B., Molhim, M., Al-Rousan, M., Dynamic model of an HVAC system for control analysis, Energy 30 (2005) 1729-1745.

Tassou, S.A., Grace, I.N., Fault diagnosis and refrigerant leak detection in vapour compression refrigeration system, International Journal of Refrigeration 28 (5) (2005) 680-688.

Torres, S.L., Sakamoto, Y., Façade design optimization for daylight with a simple genetic algorithm, In Proceedings of the Building Simulation, 2007.

Tso, K.F., Yau, K.W., Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks, Energy 32 (9) (2007) 1761-1768.

Tuhus-Dubrow, D., Krarti, M., Genetic-algorithm based approach to optimize building envelope design for residential buildings, Building and Environment 45 (7) (2010) 1574-1581.

Vahdatpour A., Amini N., Sarrafzadeh M., Towards unsupervised activity discovery

using multi-dimensional motif detection in time series, IJCAI 2009 21st International Joint Conference on Artificial Intelligence.

Vanetik N. Computing frequent graph patterns from semistructured data. ICDM 2002, 458-465.

Vega-Pons S., Ruiz-Schulcoper J., A survey of clustering ensemble algorithms, International Journal of Pattern Recognition and Artificial Intelligence 25 (2011) 337-372.

Volk, R., Stengel, J., Schultmann, F., Building information modeling (BIM) for existing buildings – literature review and future needs, Automation in Construction 38 (2014) 109-127.

Wang, S.W., Cui, J.T., Sensor FDD and estimation in centrifugal chiller systems using principal component analysis method, Applied Energy 82 (3) (2005) 197-213.

Wang, S.W., Ma, Z.J., Supervisory and optimal control of building HVAC systems: A review, HVAC&R Research 14 (1) (2008) 3-32.

Wang, S.W., Ma, Z.J., Supervisory and optimal control of building HVAC systems: A review, HVAC&R Research 14 (1) (2008) 3-32.

Wang, S.W., Ma, Z.J., Supervisory and optimal control of building HVAC systems: A review, HVAC&R Research 14 (1) (2008) 3-32.

Wang, S.W., Sun, Y.J., Ma, Z.J., Online optimal control strategy for multiple-chiller systems, Journal Chemical Engineering (CIESC) 62 (S2) (2010) 86-92.

Wang, S.W., Xiao, F., AHU sensor fault diagnosis using principal component analysis, Energy and Buildings 36 (2) (2004) 147-160.

Wang, S.W., Xiao, F., Detection and diagnosis of AHU sensor faults using principal component analysis method, Energy Conversion and Management 45 (17) (2004) 2667-2686.

Wang, S.W., Zhou, Q., Xiao, F., A system-level fault detection and diagnosis strategy

for HVAC systems involving sensor faults, Energy and Buildings 42 (4) (2010) 477-490.

Wang, W.M., Zmeureanu, R., Rivard, H., Applying multi-objective genetic algorithms in green building design optimization, Building and Environment 40 (11) (2005) 1512-1525.

Washio T., Motoda H. State of the art of graph-based data mining. SIGKDD Explorations: Newsletter of the ACM Special Interest Group on Knowledge Discovery & Data Mining 5 (1) (2003) 59-68.

Wen, J., Li, S., Application of pattern matching method for detecting faults in air handling unit system, Automation in Construction 43 (2014) 49-58.

West, S.R., Ward, J.K., Wall, J., Trial results from a model predictive control and optimization system for commercial building HVAC, Energy and Buildings 72 (2014) 271-279.

Williams, T.P., Gong, J., Predicting construction cost overruns using text mining, numerical data and ensemble classifiers, Automation in Construction 43 (1) (2014) 23-29.

Worlein M., Meinl T., Fisher I., Philippsen M. A quantitative comparison of the subgraph miners MoFa, gSpan, FFSM and Gaston. Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2005, 392-404.

Wu, S.M., Clements-Croome, D., Understanding the indoor environment through mining sensory data – A case study, Energy and Buildings 39 (11) (2007) 1183-1191.

Wu, S.Y., Sun, J.Q., Cross-level fault detection and diagnosis of building HVAC systems, Building and Environment 46 (8) (2011) 1558-1566.

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

Xiao, F., Wang, S.W., Process and methodologies of lifecycle commissioning of

HVAC systems to enhance building sustainability, Renewable and Sustainable Energy Reviews 13 (5) (2009) 1144-1149.

Xiao, F., Wang, S.W., Progress and methodologies of lifecycle commissioning for HVAC systems to enhance building sustainability-a review, Renewable and Sustainable Energy Reviews 13 (5) (2009) 1144-1149.

Xiao, F., Wang, S.W., Xu, X.H., Ge, G.M., An isolation enhanced PCA method with expert-based multivariate decoupling for sensor FDD in air-conditioning systems, Applied Thermal Engineering 29 (4) (2009) 712-722.

Xiao, F., Zhao, Y., Wen, J., Wang, S.W., Bayesian network based FDD strategy for variable air volume terminals, Automation in Construction 41 (2014) 106-118.

Xiao, F., Zhao, Y., Wen, J., Wang, S.W., Bayesian network based FDD strategy for variable air volume terminals, Automation in Construction 41 (13) (2014) 106-118.

Xue X., Wang S.W., Sun Y.J., Xiao F., An interactive building power demand management strategy for facilitating smart grid optimization, Applied Energy 116 (2014) 297-310.

Yan X., Han J.W. gSpan: Graph-based substructure pattern mining. Proceedings of International Conference on Data Mining, 2002, 721-724.

Yan X.F., Han J.W. CloseGraph: Mining closed frequent graph patterns. The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August, 2003.

Yan, K., Shen, W., Mulumba, T., Afshari, A., ARX model based fault detection and diagnosis for chillers using support vector machines, Energy and Buildings 81 (2014) 287-295.

Yang, J., Rivard, H., Zmeureanu, R., On-line building energy prediction using adaptive artificial neural networks, Energy and Buildings 37 (12) (2005) 1250-1259.

Yu Z., Haghighat F., Fung C.M., Zhou L., A novel methodology for knowledge

discovery through mining associations between building operational data, Energy and Buildings 47 (50) (2012) 430-440.

Yu, W.D., Lin, H.W., A VaFALCON neuro-fuzzy system for mining of imcomplete construction databases, Automation in Construction 15 (1) (2006) 20-32.

Yu, W.D., Liu, Y.C., Hybridization of CBR and numeric soft computing techniques for mining of scarce construction databases, Automation in Construction 15 (1) (2006) 33-46.

Yu, Y.B., Woradechjumroen, D., Yu, D.H., A review of fault detection and diagnosis methodologies on air-handling units, Energy and Buildings 82 (2014) 550-562.

Yu, Z., Haghighat, F., Fung, C.M., Yoshino, H., A decision tree method for building energy demand modeling, Energy and Buildings 42 (10) (2010) 1637-1646.

Yun, K., Luck, R., Mago, P.J., Cho, H., Building hourly thermal load prediction using an indexed ARX model, Energy and Buildings 54 (2012) 225-233.

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

Zhao, H.X., Magoules, F., A review on the prediction of building energy consumption, Renewable and Sustainable Energy Reviews 16 (6) (2012) 3586-3592.

Zhao, H.X., Magoules, F., A review on the prediction of building energy consumption, Renewable and Sustainable Energy Reviews 16 (6) (2012) 3586-3592.

Zhao, H.X., Magoules, F., Feature selection for predicting building energy consumption based on statistical learning method, Journal of Algorithms & Computational Technology 6 (1) (2012) 59-77.

Zhao, H.X., Magoules, F., Parallel support vector machines applied to the prediction of multiple building energy consumption, Journal of Algorithms & Computational Technology 4 (2) (2010) 231-250.

Zhao, Y., Wang, S.W., Xiao, F., A statistical fault detection and diagnosis method for centrifugal chillers based on exponentially-weighted moving average control charts and support vector regression, Applied Thermal Engineering 51 (1-2)

(2013) 560-572.

Zhao, Y., Wang, S.W., Xiao, F., A system-level incipient fault detection method for HVAC systems, HVAC&R Research 19 (3) (2013) 593-601.

Zhao,Y., Xiao, F., Wang, S.W., An intelligent chiller fault detection and diagnosis methodology using Bayesian belief network, Energy and Buildings 57 (2013) 278-288.

Zhou, Q., Wang, S.W., Xiao, F., A novel strategy for the fault detection and diagnosis of centrifugal chiller systems, HVAC&R Research 15 (1) (2009) 57-75.

APPENDIX A - HighDimOut Package

The HighDimOut package is developed by Fan Cheng in *R* to facilitate the task of outlier detection in high-dimensional datasets. The package can be downloaded from the Comprehensive R Archive Network (CRAN) https://cran.r-project.org. The technical description and examples can be found in the package manual. Three main algorithms are implemented, including the angle-based outlier detection (ABOD), subspace outlier detection (SOD), and feature-bagging outlier detection (FBOD). The package codes are listed below.

****

#' A function to calculate the shared nearest neighbors (SNN)

#' SNN is reported to be more robust than k nearest neighbors.

#' Firstly, the k nearest neighbor distances for each observation is calculated.

#' Then, the shared nearest neighbor similarity is calculated based on the result of k nearest neighbor.

#' Note that k.nn should be greater than k.sel.

#' @import plyr

#' @param data is the data frame containing the observations (should be numeric data). Each row represents an observation and each variable is stored in one column.

#' @param k.nn specifies the value used for calculating the shared nearest neighbors.

#' @param k.sel specifies the number of shared nearest neighbors

#' @return The function returns the matrix containing the indices of top k shared nearest neighbors for each observation

#' @examples

#' Func.SNN(data=TestData[,1:2], k.nn=5, k.sel=3)

#' @export

Func.SNN <- function(data, k.nn, k.sel) {

```
        #Get the knn index
        mat.ind <- FNN::get.knn(data = data, k = k.nn)$nn.index
        #Define distance function
        func.dist <- function(x1, x2) {
                length(intersect(x1, x2))
        }
        #Count the distance using the customized function
        mat.count <- as.matrix(proxy::dist(x = mat.ind, method = func.dist, diag = T,
upper = T))
        #Formulate the final matrix for use
        mat.final <- plyr::aaply(.data = mat.count, .margins = 1, .fun = function(x)
{order(x, decreasing=T)[1:k.sel]})
        return(mat.final)
}
**********************************************************************
#' Angle-based outlier detection (ABOD) algorithm
#' This function performs the basic and aprroximated version of angle-based outlier
detection algorithm.
#' The ABOD method is especially useful for high-dimensional data, as angle is a
more robust measure than distance in high-dimensional space.
#' The basic version calculate the angle variance based on the whole data. The results
obtained are more reliable. However, the speed can be very slow.
#' The approximated version calculate the angle variance based on a subset of data
and thereby, increasing the calculation speed.
#' This function is based on the work of Krigel, H.P., Schubert, M., Zimek, A.,
Angle-based outlier detection in high dimensional data, 2008.
#' @import foreach
#' @import plyr
#' @import ggplot2
```

```
#' @param data is the data frame containing the observations. Each row represents an observation and each variable is stored in one column.
#' @param basic is a logical value, indicating whether the basic method is used. The speed of basic version can be very slow if the data size is large.
#' @param perc defines the percentage of data to use when calculating the angle variance. It is only needed when basic=F.
#' @return The function returns the vector containing the angle variance for each observation
#' @examples
#' library(ggplot2)
#' res.ABOD <- Func.ABOD(data=TestData[,1:2], basic=FALSE, perc=0.2)
#' data.temp <- TestData[,1:2]
#' data.temp$Ind <- NA
#' data.temp[order(res.ABOD, decreasing = FALSE)[1:10],"Ind"] <- "Outlier"
#' data.temp[is.na(data.temp$Ind),"Ind"] <- "Inlier"
#' data.temp$Ind <- factor(data.temp$Ind)
#' ggplot(data = data.temp) + geom_point(aes(x = x, y = y, color=Ind, shape=Ind))
#' @export
Func.ABOD <- function(data, basic=FALSE, perc) {
    i=j=NULL
    if(basic==T) {
        res <- foreach::foreach(i = 1:(dim(data)[1]), .combine = c) %dopar% {
            obs <- data[i,]
            com <- t(combn(x = c(1:dim(data)[1])[-i], m = 2))

            cos.angles <- foreach(j = 1:(dim(com)[1]), .combine = c) %do% {
                vec.1 <- data[com[j,1],] - obs
                vec.2 <- data[com[j,2],] - obs
```

```
                    round(acos(sum(vec.1 *
vec.2)/(sqrt(sum(vec.1^2))*sqrt(sum(vec.2^2)+0.01)))/(sqrt(sum(vec.1^2))*sqrt(sum(
vec.2^2)+0.01)), digits = 2)
                }
                return(var(x = cos.angles))
            }
            return(res)
        } else {
            nu <- round(dim(data)[1]*perc, digits = 0)
            res <- foreach::foreach(i = 1:(dim(data)[1]), .combine = c) %dopar% {
                obs <- data[i,]
                index.used <- sample(x = c(1:dim(data)[1])[-i], size = nu, replace = F)
                com <- t(combn(x = index.used, m = 2))

                cos.angles <- foreach(j = 1:(dim(com)[1]), .combine = c) %do% {
                    vec.1 <- data[com[j,1],] - obs
                    vec.2 <- data[com[j,2],] - obs
                    round(acos(sum(vec.1 *
vec.2)/(sqrt(sum(vec.1^2))*sqrt(sum(vec.2^2)+0.01)))/(sqrt(sum(vec.1^2))*sqrt(sum(
vec.2^2)+0.01)), digits = 2)
                }
                return(var(x = cos.angles))
            }
            return(res)
        }
}
**********************************************************************
#' Subspace outlier detection (SOD) algorithm
#' This function performs suspace outlier detection algorithm
```

```
#' The implemented method is based on the work of Krigel, H.P., Kroger, P., Schubert, E., Zimek, A., Outlier detection in axis-parallel subspaces of high dimensional data, 2009.
#' @import foreach
#' @import plyr
#' @import ggplot2
#' @param data is the data frame containing the observations. Each row represents an observation and each variable is stored in one column.
#' @param k.nn specifies the value used for calculating the shared nearest neighbors. Note that k.nn should be greater than k.sel.
#' @param k.sel specifies the number shared nearest neighbors. It can be interpreted as the number of reference set for constructing the subspace hyperplane.
#' @param alpha specifies the lower limit for selecting subspace. 0.8 is set as default as suggested in the original paper.
#' @return The function returns a vector containing the SOD outlier scores for each observation
#' @examples
#' library(ggplot2)
#' res.SOD <- Func.SOD(data = TestData[,1:2], k.nn = 10, k.sel = 5, alpha = 0.8)
#' data.temp <- TestData[,1:2]
#' data.temp$Ind <- NA
#' data.temp[order(res.SOD, decreasing = TRUE)[1:10],"Ind"] <- "Outlier"
#' data.temp[is.na(data.temp$Ind),"Ind"] <- "Inlier"
#' data.temp$Ind <- factor(data.temp$Ind)
#' ggplot(data = data.temp) + geom_point(aes(x = x, y = y, color=Ind, shape=Ind))
#' @export
Func.SOD <- function(data, k.nn, k.sel, alpha=0.8) {
    i=j=NULL
    mat.ref <- Func.SNN(data = data, k.nn = k.nn, k.sel = k.sel)
    res <- foreach::foreach(i=1:dim(data)[1], .combine = c) %dopar% {
```

```
            obs <- data[i,]
            ref <- as.matrix(data[mat.ref[i,],])
            means <- colMeans(ref)
            var.total <- sum(aaply(.data = ref, .margins = 1, .fun = function(x)
sum((x-means)^2)))/k.sel
            var.expect <- alpha*var.total/dim(data)[2]
            var.actual <- foreach(j = 1:dim(ref)[2], .combine = c) %dopar% {
                  var(ref[,j])
            }
            var.ind <- ifelse(var.actual<var.expect, yes = 1, no = 0)
            res.hyper <- sqrt(sum(var.ind*(obs-means)^2))/length(which(var.ind==1))
            return(res.hyper)
      }
      return(res)
}
**********************************************************************
#' Feature-bagging outlier detection (FBOD) algorithm
#' This function performs feature-bagging based outlier detection algorithm
#' The implemented method is based on the work of "Lazarevic, A., Kumar, V.,
Feature bagging for outlier detection, 2005"
#' This method can be regarded as an ensemble method, which based on the results of
local outlier factor (LOF).
#' During each iteration, a random subset of variables, whose size is randomly chosen
between d/2 to d (where d is the dimensionality of the input data), is selected.
#' The LOF method is applied to calculate the LOF scores based on the selected data
subset.
#' The final score of FBOD is the cumulative sum of each iteration.
#' @import foreach
#' @import plyr
#' @import ggplot2
```

```
#' @param data is the data frame containing the observations. Each row represents an
observation and each variable is stored in one column.
#' @param iter is the iteration used.
#' @param k.nn is the value used for calculating the LOF score
#' @return The function returns a vector containing the FBOD outlier scores for each
observation
#' @examples
#' library(ggplot2)
#' res.FBOD <- Func.FBOD(data = TestData[,1:2], iter=10, k.nn=5)
#' data.temp <- TestData[,1:2]
#' data.temp$Ind <- NA
#' data.temp[order(res.FBOD, decreasing = TRUE)[1:10],"Ind"] <- "Outlier"
#' data.temp[is.na(data.temp$Ind),"Ind"] <- "Inlier"
#' data.temp$Ind <- factor(data.temp$Ind)
#' ggplot(data = data.temp) + geom_point(aes(x = x, y = y, color=Ind, shape=Ind))
#' @export
Func.FBOD <- function(data, iter, k.nn) {
      i=NULL
      res <- foreach(i=1:iter, .combine = cbind) %dopar% {
            d <- dim(data)[2]
            l <- sample(x = round(d/2, digits = 0):(d-1), size = 1)
            ind <- sample(x = 1:d, size = l, replace = F)
            data.use <- data[,ind]
            score <- DMwR::lofactor(data = data.use, k = k.nn)
            return(score)
      }
      res[is.nan(res)] <- NA
      res[is.infinite(res)] <- max(res[!is.infinite(res)], na.rm = T)
      res.final <- plyr::aaply(.data = res, .margins = 1, .fun = function(x) mean(x,
na.rm = T))
```

```
    return(round(res.final, digits = 3))
}
**********************************************************************
#' Outlier score transformation
#' This function calculate the transformed outlier scores, with the aim of unifying the results from different methods.
#' The method is based on the work of Kriegel, H.P., Kroger, P., Schubert, E., Zimek, A., Interpreting and unifying outlier scores, 2011.
#' It consists of two steps, regularization and normalization.
#' For the ABOD scores, logarithmic inversion is used for regularization
#' For the SOD scores, no action is taken to perform regularization
#' For the FBOD method, the basic regularization, i.e., score-1, is used for regularization
#' For the normalization step, the gaussian scaling method is used.
#' The final output can be interpreted as the outlier probability, ranging from 0 to 1.
#' @param raw.score is the scores returned by each method
#' @param method should be a character specifying    the method used to generate the raw score. It has 3 possible values, "ABOD", "SOD", and "FBOD".
#' @return The function returns the transformed outlier scores
#' @export
Func.trans <- function(raw.score, method) {
    #Regularization
    if(method=="FBOD") {score.reg <- raw.score-1}
    if(method=="ABOD") {score.reg <- -log(x = raw.score/max(raw.score), base = 10)}
    if(method=="SOD") {score.reg <- raw.score}

    #Normalization
    erf <- function(x) 2*pnorm(x*sqrt(2))-1
    if (sd(score.reg, na.rm = T)==0) {score.norm <- rep(0, length(raw.score))} else {
```

```
            score.norm <- erf(x=(score.reg-mean(score.reg, na.rm =
T))/(sqrt(x=2)*sd(score.reg, na.rm = T)))
            score.norm[which(score.norm<0)] <- 0
        }
        return(score.norm)
}
```

****

APPENDIX B - TSMining Package

The TSMining *R* package is developed by Fan Cheng in *R* to facilitate the task of temporal knowledge discovery. The package can be downloaded from the Comprehensive R Archive Network (CRAN). It mainly implements the symbolic approximation aggregation (SAX), the univariate and multivariate motif discovery, and some *ggplot2* based visualization methods. The technical description and examples can be found in the package manual. The package codes are listed below.

```
**********************************************************************
#' A function to perform symbolic approximation aggregate (SAX) for time series data
#' The function create SAX symbols for a univariate time series. The details of this method can be referred to J. Lin, E. Keogh, L. Wei, S. Lonardi. Experiencing SAX: a novel symbolic representation of time series
#' @import foreach
#' @import plyr
#' @param x is a numeric vector representing the univariate time series
#' @param w is the word size and should be an integer
#' @param a is the alphabet size and should be an integer
#' @param eps is the minimum threshold for variance in x and should be a numeric value. If x has a smaller variance than eps, it will represented as a word using the middle alphabet.
#' @param norm is a logical value deciding whether standardization should be applied to x. If True, x is standardized using mean and standard deviation
#' @return The function returns a SAX representation of x
#' @examples
#' x <- runif(n = 20, min = 0, max = 20)
#' Func.SAX(x = x, w = 5, a = 5, eps = .01, norm = TRUE)
#' @export
Func.SAX    <- function(x, w, a, eps, norm) {
      i=NULL
      if(sd(x) <= eps) {sym <- rep(letters[round((1+a)/2, digits = 0)], w)} else {
            #Normalize the data to have 0 mean and 1 standard deviation before piecewise aggregation
            if(norm==TRUE) {data.nor <- (x-mean(x))/sd(x)} else {
                  data.nor <- x
```

```
        }

        #Perform the piecewise aggregation
        ind <- round(seq(from = 1, to = length(data.nor), length.out = w+1), digits =
0)
        pieces <- foreach::foreach(i=1:(length(ind)-1), .combine = c) %do% {
            if(i!=(length(ind)-1)) {piece <- data.nor[ind[i]:(ind[i+1]-1)]} else
{piece <- data.nor[ind[i]:ind[i+1]]}
            return(mean(piece, na.rm = T))
        }

        #Perform alphabet assignment
        let <- letters[1:a]
        #Create breaks points based on Gaussian normal distribution
        bks <- round(qnorm(p = seq(from = 0, to = 1, length.out = a+1)), digits = 2)
        sym <- foreach::foreach(i=1:length(pieces), .combine = c) %do% {
            obs <- pieces[i]
            let[max(which(bks<obs))]
        }
    }

    return(sym)
}
**************************************************************************
#' A function to create the distance matrix for alphabets
#' This function create a distance matrix for alphabets used for SAX transformation
#' @param a is an integer specifying the alphabet size.
#' @return The function returns a matrix showing the distance between alphabets
#' @examples
#' Func.matrix(a=5)
#' @export
Func.matrix <- function(a) {
    i=j=NULL
    let <- letters[1:a]
    bks <- round(qnorm(p = seq(from = 0, to = 1, length.out = (a+1))), digits = 2)
    bks.upd <- bks[-c(1,length(bks))]

    #Create the matrix for distance calculation
    dist.m <- matrix(data = NA, nrow = a, ncol = a, dimnames = list(let, let))
    for(i in 1:dim(dist.m)[1]) {
        for(j in 1:dim(dist.m)[2]) {
            dist.m[i,j] <- ifelse(abs(i-j)<=1, 0,
bks.upd[max(c(i,j))-1]-bks.upd[min(c(i,j))])
        }
```

```
    }
    return(dist.m)
}
**********************************************************************
#' A function to calculate the distance between two SAX representations
#' This function calculates the distance between two SAX representations
#' @import foreach
#' @import plyr
#' @param x is a SAX representations.
#' @param y is a SAX representations. It should have the same length as x.
#' @param mat is the distance matrix created by \code{Func.matrix}
#' @param n is the length of the original time series before the SAX transformation
#' @return The function returns a numeric value, which is the distance between two SAX representations
#' @examples
#' #Assuming the original time series has a length of 20, n=20
#' #Assuming the time series is transformed into SAX representations using w=4 and a=4
#' #Assuming one is a,b,c,d and the other is d,b,c,d
#' Func.dist(x=c("a","b","c","d"), y=c("d","b","c","d"), mat=Func.matrix(a=4), n=20)
#' @export
Func.dist <- function(x, y, mat, n) {
    i=NULL
    w <- length(x)
    d <- foreach::foreach(i=1:length(x), .combine = c) %do% {
        mat[which(rownames(mat)==x[i]), which(colnames(mat)==y[i])]
    }
    return(sqrt(sum(d^2))*sqrt(n/w))
}
**********************************************************************
#' A function implementing the univariate motif discovery algorithm using random projection
#' The function implements the univariate motif discovery algorithm proposed in B. Chiu, E. Keogh, S. Lonardi. Probabilistic discovery of time series motifs. ACM SIGKDD, Washington, DC, USA, 2003, pp. 493-498.
#' @import foreach
#' @import plyr
#' @param ts is a numeric vector representing the univarate time series
#' @param global.norm is a logical value specifying whether global standardization should be used for the whole time series
#' @param local.norm is a logical value specifying whether local standardization should be used for each subsequences
#' @param window.size is a integer which defines the length of the sliding window used to create subsequences
```

#' @param overlap is a numeric value ranging from 0 to 1. It defines the percentage of overlapping when using sliding window to create subsquences. 0 means subsequences are created without overlaps. 1 means subsequences are created with the maximum overlap possible.
#' @param w is an integer which defines the word size used for SAX transformation
#' @param a is an integer which defines the alphabet size used for SAX transformation
#' @param eps is the minimum threshold for variance in subsequence and should be a numeric value. If the subsequence considered has a smaller variance than eps, it will be represented as a word using the middle alphabet. The default value is 0.1
#' @param mask.size is the mask size used for random projection. It should be an integer ranging from 1 to the word size w
#' @param iter is an integer which specifies the iteration number in random projection, default value is 25
#' @param max.dist.ratio is a numeric value used to add other possible members to a motif candidate. Default value is 1.2. Each motif candidate has two subsequences. The distance between these two candidates are calculated as a baseline, denoted as BASE. Any subsequence, whose distance to the motif candidate is smaller than max.dist.ratio*BASE, is considered as a member of that motif candidate.
#' @param count.ratio.1 defines the ratio between the iteration number and the minimum value in the collision matrix to be considered as motif candidate. Default value is 1.5. For instance, if the iter is 100, any pair of subsequence, which results in a value larger than 67 in the collision matrix, is considered as a motif candidate.
#' @param count.ratio.2 defines the ratio between the maximum counts in the collision matrix and any other count values that will be considered as potential members to a motif candidate
#' @return The function returns a list of 6 elements. The first element is Subs, which is a data frame containing all the subsequences in original data formatThe second element is Subs.SAX, which is a data frame containing all the subsequences in SAX representations. The third element is Motif.raw, which is a list showing the motifs discovered in original data format.The fourth element is Motif.SAX, which is a list showing the motifs discovered in SAX representations. The fifth element is Collision.matrix, which is matrix containing the results of random projection. The sixth element is Indices, which is a list showing the starting positions of subsequences for each motif discovered.
#' @examples
#' #Perform the motif discovery for the first time series in the example data
#' data(test)
#' res.1 <- Func.motif(ts = test$TS1, global.norm = TRUE, local.norm = FALSE,
#' window.size = 10, overlap = 0, w = 5, a = 3, mask.size = 3, eps = .01)
#' #Check the number of motifs discovered
#' length(res.1$Indices)
#' #Check the starting positions of subsequences of each motif discovered
#' res.1$Indices

```
#' @export
Func.motif <- function(ts, global.norm, local.norm, window.size, overlap, w, a, mask.size,
                       eps=0.1, iter=25, max.dist.ratio=1.2, count.ratio.1=1.5, count.ratio.2=1.2) {
    i=j=q=m=k=l=g=h=u=NULL
    #Perform uniform normalization
    if(global.norm==TRUE) {ts.nor <- (ts - mean(ts))/sd(ts)} else {ts.nor <- ts}

    #Create the subsequence data frame
    b <- ifelse(overlap==1, yes = 1, no = round((1-overlap)*window.size, digits = 0))
    ts.subs <- foreach::foreach(i=seq(from = 1, to = length(ts), by = b), .combine = rbind) %do% {
        c(i,subs.temp <- ts.nor[i:(i+window.size-1)])
    }
    ts.subs <- na.omit(ts.subs)

    #Local normalization if needed
    ts.sax <- foreach::foreach(i=1:dim(ts.subs)[1], .combine = rbind) %do% {
        if (sd(ts.subs[i,-1])<=eps) {sax.temp <- rep(letters[round(a+1/2)], times = w)} else {
            sax.temp <- Func.SAX(x = ts.subs[i,-1], w = w, a = a, eps = eps, norm = local.norm)
        }
        c(ts.subs[i,1], sax.temp)
    }

    ts.sax <- as.data.frame(ts.sax, stringsAsFactors = FALSE)
    colnames(ts.sax) <- c("StartP", 1:w)
    ts.sax$StartP <- as.numeric(ts.sax$StartP)

    #Perform the random projection
    col.mat <- matrix(data = 0, nrow = dim(ts.sax)[1], ncol = dim(ts.sax)[1])
    for(i in 1:iter) {
        col.pos <- sort(sample(x = 2:dim(ts.sax)[2], size = mask.size, replace = F), decreasing = F)
        sax.mask <- ts.sax[,col.pos]

        unique.lab <- unique(sax.mask)

        mat <- foreach::foreach(j = 1:dim(unique.lab)[1], .combine = rbind) %do% {
```

```
            indices <- foreach::foreach(q = 1:dim(sax.mask)[1], .combine =
c) %do% {
                identical(as.character(sax.mask[q,]), as.character(unique.lab[j,]))
            }
        }

        for(m in 1:dim(mat)[1]) {
            if(length(which(mat[m,]==TRUE))>1) {
                com <- t(combn(x = which(mat[m,]==TRUE), m = 2))
                col.mat[com] <- col.mat[com] + 1
            }
        }
    }

    #Extact the tentative motif pair
    counts <- sort(col.mat, decreasing=TRUE)
    counts.sel <- counts[which(counts>=(iter/count.ratio.1))]

    motif.pair <- foreach::foreach(k = 1:length(unique(counts.sel)), .combine =
rbind) %do% {
        arrayInd(which(col.mat==unique(counts.sel)[k]), .dim = dim(col.mat))
    }

    indices <- foreach::foreach(l = 1:dim(motif.pair)[1]) %do% {
        pair <- c(ts.sax[motif.pair[l,1],1], ts.sax[motif.pair[l,2],1])

        cand.1 <- ts.subs[motif.pair[l,1],-1]
        cand.2 <- ts.subs[motif.pair[l,2],-1]
        dist.raw <- sqrt(sum((cand.1 - cand.2)^2))

        col.no <- col.mat[motif.pair[l,1],]
        ind.cand <- which(col.no > (max(col.no)/count.ratio.2))
        if(length(ind.cand)>1) {
            ind.temp <- ind.cand[-which(ind.cand == motif.pair[l,2])]
            if(length(ind.temp)==1) {
                df.cand.sel <- as.matrix(ts.subs[ind.temp,-1])
                dist.res <- plyr::aaply(.data = df.cand.sel, .margins = 2, .fun =
function(x) sqrt(sum((cand.1-x)^2)))
                ind.final                                                    <-
ts.sax[ind.temp[which(dist.res<=max.dist.ratio*dist.raw)],1]
            } else {
                df.cand.sel <- ts.subs[ind.temp,-1]
                dist.res <- plyr::aaply(.data = df.cand.sel, .margins = 1, .fun =
function(x) sqrt(sum((cand.1-x)^2)))
```

```
                    ind.final <-
ts.sax[ind.temp[which(dist.res<=max.dist.ratio*dist.raw)],1]
                }} else {
                    ind.final <- NULL
                }

        pair.final <- c(pair, ind.final)
    }

    #Combine the indices if there is any overlap
    vec.subset <- rep(0, length(indices))
    foreach::foreach(g = 1:(length(indices)-1), .combine = rbind) %do% {
        for (h in (g+1):length(indices)) {
            if(length(which(indices[[g]] %in% indices[[h]]))>0) {
                indices[[h]] <- unique(c(indices[[g]], indices[[h]]))
                vec.subset[g] <- 1
            }
        }
    }
    indices <- indices[vec.subset==0]

    motif.raw <- foreach::foreach(u = 1:length(indices)) %do% {
        ts.subs[which(ts.subs[,1] %in% indices[[u]]),]
    }

    motif.sax <- foreach::foreach(u = 1:length(indices)) %do% {
        ts.sax[which(ts.sax[,1] %in% indices[[u]]),]
    }

    return(list(Subs=ts.subs, Subs.SAX=ts.sax, Motif.raw=motif.raw,
Motif.SAX=motif.sax, Collision.matrix=col.mat, Indices=indices))
}
**********************************************************************
#' A function to implement the multivariate motif discovery
#' This function implements the multivariate motif discovery method proposed in A. Vahdatpour, N. Amini, M. Sarrafzadeh. Towards unsupervised activity discovery using multi-dimensional motif detection in time series. IJCAI 2009 21st International Joint Conference on Artificial Intelligence.
#' @import foreach
#' @import plyr
#' @param motif.list is a list of lists, each contains the univariate motifs discovered in a univariate time series. The component of motif.list is the results of Func.motif()$Indices, which store the starting position of subsequences of each univariate motif
```

```
#' @param window.sizes is a vector containing the length of motifs in each univariate time series. It should have the same order as components in motif.list.
#' @param alpha is a numeric ranging from 0 to 1. It specifies the minimum correlation between two univariate motifs before considered as multivariate motifs
#' @return The function returns a list containing two elements. The first element is Motif, which is a list containing the univarate motif IDs for different multivariate motifs. e.g., if there are two univariate time series and each has 3 motifs, then univariate ID is from 1 to 6. The second element is Info, which is a list storing the information of univariate motifs for different multivariate motifs
#' @examples
#' data(test)
#' #Perform univariate motif discovery for each dimension in the example data
#' res.1 <- Func.motif(ts = test$TS1, global.norm = TRUE, local.norm = FALSE,
#' window.size = 10, overlap = 0, w = 5, a = 3, mask.size = 3, eps = .01)
#' res.2 <- Func.motif(ts = test$TS2, global.norm = TRUE, local.norm = FALSE,
#' window.size = 20, overlap = 0, w = 5, a = 3, mask.size = 3, eps = .01)
#' #Perform multivaraite motif discovery
#' res.multi <- Func.motif.multivariate(motif.list = list(res.1$Indices, res.2$Indices),
#' window.sizes = c(10,20), alpha = .8)
#' @export
**********************************************************************
Func.motif.multivariate <- function(motif.list, window.sizes, alpha) {
    i=j=q=p=t=o=f=x=z=NULL
    #Get the total motif.no
    tot.no <- sum(plyr::laply(.data = motif.list, .fun = length))

    #Create the weight matrix
    w.mat <- matrix(data = 0, nrow = tot.no, ncol = tot.no)

    #Get the characteristics of motifs
    info <- foreach::foreach(i=1:length(motif.list), .combine = rbind) %do% {
        info.sub <- foreach::foreach(j=1:length(motif.list[[i]]), .combine =
rbind) %do% {
            info.sub.sub <-
foreach::foreach(q=1:length(motif.list[[i]][[j]]), .combine = rbind) %do% {
                c(i,j,q, motif.list[[i]][[j]][q],
motif.list[[i]][[j]][q]+window.sizes[i]-1)
            }
        }
    }
    rownames(info) <- 1:dim(info)[1]
    info <- as.data.frame(info)
    colnames(info) <- c("Variable","Motif","Member","StartP","EndP")
```

```
info$Lab <- as.numeric(as.character(factor(info$Variable*100 + info$Motif, levels = unique(info$Variable*100 + info$Motif), labels = 1:length(unique(info$Variable*100 + info$Motif)))))
info.ori <- info

#Generate the weights
pb <- txtProgressBar(min = 0, max = length(unique(info$Lab)))
temp <- foreach::foreach (i = 1:length(unique(info$Lab))) %do% {
    setTxtProgressBar(pb = pb, value = i)
    mot.con <- info[which(info$Lab==i),]
    lab.con <- unique(mot.con$Lab)
    n <- dim(mot.con)[1]
    variable.con <- unique(mot.con$Variable)
    mot.com.all <- info[-which(info$Variable==variable.con),]

    temp.ind <- foreach::foreach (j = 1:length(unique(mot.com.all$Lab)), .combine = rbind) %do% {
        mot.com <- mot.com.all[which(mot.com.all$Lab==unique(mot.com.all$Lab)[j]),]
        lab.com <- unique(mot.com$Lab)
        count <- 0
        temp.ind.ind <- foreach::foreach (p = 1:dim(mot.con)[1], .combine = c) %do% {
            res.temp <- foreach::foreach (x = 1:dim(mot.com)[1], .combine = c) %do% {
                out <- ifelse(mot.con[p,"StartP"]>mot.com[x,"EndP"] | mot.con[p,"EndP"]<mot.com[x,"StartP"], yes = 0, no = 1)
                return(out)
            }
            if(sum(res.temp)!=0) {count <- count+1}
            return(ifelse(sum(res.temp)!=0, yes = 1, no = 0))
        }
        w.mat[lab.con ,lab.com] <- round(count/n, 2)
        return(c(lab.con, lab.com, temp.ind.ind))
    }
    return(temp.ind)
}
w.ori <- w.mat

#Perform grouping
#Get the occurrence of each motifs
occurence <- plyr::daply(.data = info, .variables = "Lab", .fun = function(x) dim(x)[1])
ord <- order(occurence, decreasing = T)
```

```
      re.temp <- foreach::foreach(t = 1:length(ord)) %do% {
            w <- w.mat[ord[t],]
            ind.add <- which(w>alpha)
            s <- c(ord[t], ind.add)

            #Remove some of the occurences of motif j when included by motif i
            if(length(ind.add)>0) {
                  out <- foreach::foreach(o = 1:length(ind.add), .combine = c) %do% {
                        obs.con <- info[which(info$Lab==ind.add[o]),]
                        mat <- temp[[ind.add[o]]]
                        w.mat[ind.add[o], ord[t]] <- 0
                        oo <- which(info$Lab==ind.add[o] & info$Member %in%
c(which(mat[which(mat[,2]==ord[t]),-c(1,2)]==1)))
                        return(oo)
                  }
                  info <- info[-unique(out),]

                  #Update the weights associated with motif j
                  #Generate new the weights for the jth row
                  for (f in 1:length(ind.add)) {
                        if(length(which(info$Lab==ind.add[f]))!=0) {
                              mot.con <- info[which(info$Lab==ind.add[f]),]
                              lab.con <- unique(mot.con$Lab)
                              n <- dim(mot.con)[1]
                              variable.con <- unique(mot.con$Variable)
                              mot.com.all <- info[-which(info$Variable==variable.con),]

                              for (j in 1:length(unique(mot.com.all$Lab))) {
                                    mot.com                                    <-
mot.com.all[which(mot.com.all$Lab==unique(mot.com.all$Lab)[j]),]
                                    lab.com <- unique(mot.com$Lab)
                                    count <- 0
                                    for (p in 1:dim(mot.con)[1]) {
                                          res.temp      <-      foreach::foreach      (x      =
1:dim(mot.com)[1], .combine = c) %do% {
                                                out                                        <-
ifelse(mot.con[p,"StartP"]>mot.com[x,"EndP"]                                        |
mot.con[p,"EndP"]<mot.com[x,"StartP"], yes = 0, no = 1)
                                                return(out)
                                          }
                                          if(sum(res.temp)!=0) {count <- count+1}
                                    }
                                    w.mat[lab.con ,lab.com] <- round(count/n, 2)
```

```
                        }
                    }
                }
                return(sort(s, decreasing = F))
            }
        }
        ind.null <- which(plyr::laply(.data = re.temp, .fun = length)>0)
        mot.final <- re.temp[ind.null]

        indices <- foreach::foreach(z = 1:length(mot.final)) %do% {
            info.ori[which(info.ori$Lab %in% mot.final[[z]]),]
        }

        return(list(Motif=mot.final, Info=indices))
}
**************************************************************************
#' A function to prepare the dataset for visualizing the univaraite motifs discovered
#' This function create a data set for the use of visualizing the univariate motifs discovered
#' @import foreach
#' @import plyr
#' @import reshape2
#' @import ggplot2
#' @param single.ts is a numeric vector used to represent the univariate time series
#' @param window.size is the window size used to create subsequences. It is also the length of univariate motifs
#' @param motif.indices is the results of Func.motif()$Indices, which store the starting position of subsequences for each univariate motifs
#' @return The function returns a list of three elements. The first element is data.1, which can be used to show the whole time series with motifs identified highlighted. The second element is data.2, which can be used to visualize the members of each motif. It is a list containing data frames. Each data frame is designed to visualize the members in each motif.
#' @examples
#' data(test)
#' #Perform univariate motif discovery for the first dimension data in the example data
#' res.1 <- Func.motif(ts = test$TS1, global.norm = TRUE, local.norm = FALSE,
#' window.size = 10, overlap = 0, w = 5, a = 3, mask.size = 3, eps = .01)
#' data.vis <- Func.visual.SingleMotif(single.ts=test$TS1, window.size=10, motif.indices=res.1$Indices)
#' #To visualize general information of motifs discovered on the whole time series
#' library(ggplot2)
#' ggplot(data = data.vis$data.1) +
```

```
#'    geom_line(aes(x = 1:dim(data.vis$data.1)[1], y = X)) +
#'    geom_point(aes(x = 1:dim(data.vis$data.1)[1], y = X, color=Y))
#' #To visualize the detailed information of the 1st motif
#' ggplot(data = data.vis$data.2[[1]]) + geom_line(aes(x = Time, y = Value, linetype=Instance))
#' @export
Func.visual.SingleMotif <- function(single.ts, window.size, motif.indices) {
	i=j=NULL
	pos <- foreach::foreach(i=1:length(motif.indices)) %do% {
		pos.ind <- foreach::foreach(j=1:length(motif.indices[[i]]), .combine = c) %do% {
			motif.indices[[i]][j]:(motif.indices[[i]][j]+window.size-1)
		}
		return(pos.ind)
	}
	temp.1 <- data.frame(X=single.ts, Y="Ref", stringsAsFactors = F)
	for(m in 1:length(pos)) {
		temp.1[pos[[m]],"Y"] <- paste0("Motif.",m)
	}
	temp.1$Y <- factor(temp.1$Y)

	temp.2 <- foreach::foreach(i=1:length(motif.indices)) %do% {
		pos.ind <- foreach::foreach(j=1:length(motif.indices[[i]]), .combine = rbind) %do% {
			single.ts[motif.indices[[i]][j]:(motif.indices[[i]][j]+window.size-1)]
		}
		pos.ind <- cbind(pos.ind, ID=1:dim(pos.ind)[1])

		pos.trans <- reshape2::melt(data = pos.ind, id.vars = "ID")
		pos.trans <- pos.trans[-which(pos.trans$Var2=="ID"),]
		pos.trans$Var2 <- as.numeric(rep(1:window.size, each = dim(pos.ind)[1]))
		pos.trans$Var1 <- as.numeric(pos.trans$Var1)
		pos.trans <- pos.trans[with(pos.trans, order(Var1, decreasing=F)),]
		colnames(pos.trans) <- c("Instance","Time","Value")
		pos.trans[,1] <- factor(pos.trans[,1])

		return(pos.trans)
	}

	return(list(data.1=temp.1, data.2=temp.2))
}
**********************************************************************
#' A function to prepare the data for the visualization of multivariate motifs discovered
```

```
#' This function prepares the data used for visualizing multivariate motifs.
#' @import foreach
#' @import plyr
#' @import ggplot2
#' @param data is a data frame containing the multivariate time series data. Each column represents a time series.
#' @param multi.motifs is the result of Func.motif.multivariate
#' @param index is an integer which specifies the No. of multivariate motif to be plotted
#' @return The function returns a data frame for the ease of visualizing multivariate motif discovered
#' @examples
#' data(test)
#' #Perform univariate motif discovery
#' res.1 <- Func.motif(ts = test$TS1, global.norm = TRUE, local.norm = FALSE,
#' window.size = 10, overlap = 0, w = 5, a = 3, mask.size = 3, eps = .01)
#' res.2 <- Func.motif(ts = test$TS2, global.norm = TRUE, local.norm = FALSE,
#' window.size = 20, overlap = 0, w = 5, a = 3, mask.size = 3, eps = .01)
#' res.multi <- Func.motif.multivariate(motif.list = list(res.1$Indices, res.2$Indices),
#' window.sizes = c(10,20), alpha = .8)
#' #Use the function to prepare the data frame for visualizing the first multivariate motifs identified
#' data.multi <- Func.visual.MultiMotif(data = test, multi.motifs = res.multi, index = 1)
#' #Make the plot using ggplot2
#' library(ggplot2)
#' ggplot(data = data.multi) +
#'    geom_line(aes(x = T, y = X)) +
#'    geom_point(aes(x = T, y = X, col=Lab, shape=Lab)) + facet_grid(Facet~.)
#' @export
Func.visual.MultiMotif <- function(data, multi.motifs, index) {
      i=q=j=NULL
      data.trans <- foreach::foreach(q = 1:dim(data)[2], .combine = c) %do% {
            as.numeric(data[,q])
      }
      data.trans <- cbind.data.frame(X=data.trans, T=rep(1:dim(data)[1], times = dim(data)[2]),
                                                Facet=rep(1:dim(data)[2], each = dim(data)[1]))
      data.trans$Facet <- factor(data.trans$Facet)

      con <- multi.motifs$Info[[index]]
      subs <- foreach::foreach(i = 1:length(unique(con$Variable)), .combine = rbind.data.frame) %do% {
```

```
        con.sub <- con[which(con$Variable==unique(con$Variable)[i]),]
        xs <- foreach::foreach(j = 1:dim(con.sub)[1], .combine = c) %do% {
            c(con.sub[j,"StartP"]:con.sub[j,"EndP"])
        }

        xs.upd <- data.frame(X=xs, Lab=rep(con.sub$Lab,
each=con.sub[1,"EndP"]-con.sub[1,"StartP"]+1), Var.no=unique(con.sub$Variable))
        return(xs.upd)
    }

    subs$Index.adj <- (subs$Var.no-1)*dim(data)[1]+subs$X
    data.trans$Index.adj <- 1:dim(data.trans)[1]

    data.upd <- plyr::join(x = data.trans, y = subs[,-1])
    data.upd[is.na(data.upd$Lab),"Lab"] <- 0
    data.upd[is.na(data.upd$Var.no),"Var.no"] <-
as.numeric(as.character(data.upd[is.na(data.upd$Var.no),"Facet"]))
    data.upd$Lab <- factor(data.upd$Lab)
    data.upd$Var.no <- NULL

    return(data.upd)
}
**********************************************************************
```