2014

Rare Event Analysis of High Dimensional Building Operational Data Using Data Mining Techniques

Cheng Fan
The Hong Kong Polytechnic University, Hong Kong S.A.R. (China), 11901679r@connect.polyu.hk

Fu Xiao
The Hong Kong Polytechnic University, Hong Kong S.A.R. (China), linda.xiao@polyu.edu.hk

Shengwei Wang
The Hong Kong Polytechnic University, Hong Kong S.A.R. (China), shengwei.wang@polyu.edu.hk

Rare Event Analysis of High Dimensional Building Operational Data Using Data Mining Techniques

Cheng FAN, Fu XIAO*, Shengwei WANG

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

Email: linda.xiao@polyu.edu.hk; Tel: +852 2766 4194; Fax: +852 2765 7198

ABSTRACT

Today's building automation systems (BASs) are becoming increasingly complex. A typical BAS usually stores hundreds of sensor measurements and control signals at each time step, which produces massive high dimensional data sets. Traditional analysis methods for BAS data only focus on a small subset of the data, resulting in a huge information loss. Data mining techniques are more effective in knowledge extraction of massive data. This study develops a holistic methodology for analyzing the high dimensional BAS data using advanced data mining techniques, with the aim of identifying rare events in building operation. Rare event analysis helps to identify atypical building operating patterns, detect and diagnose faults, and eventually improve the building operational performance.

Two main challenges exist in performing rare event analysis of massive building operational data, i.e. the high data dimensionality and the complexity in building system operation. The former results that the conventional analytics, such as distance-based measures, lose their effectiveness, and the later negatively influences the robustness and reliability of the identification of rare events. The proposed method is specially designed to tackle these challenges by integrating the power of data mining techniques. It consists of four main steps, i.e., data preparation, rare event detection, rare event diagnosis, and post-mining. The methodology is adopted to analyze the BAS data of the tallest building in Hong Kong. Rare events are successfully detected and diagnosed, providing clues to enhance building operational performance.

Keywords: rare event analysis; data mining; building automation; clustering analysis; outlier detection ensembles

1. INTRODUCTION

Nowadays, buildings are responsible for 32% of total final energy consumption and 40% of primary energy consumption in most countries (IEA, 2013). These figures would be much higher for less industrial-oriented districts. For instance, the building sector in Hong Kong accounts for more than 90% of electricity use and over 60% of total greenhouse gas emission (EMSD, 2013). To cope with the ever-increasing burden of energy crisis, and its associated economic and environmental consequences, advanced technologies have been employed to improve building energy efficiency, such as building automation system (BAS), which integrates technologies from information science, computing science, and control theory etc. BAS enables buildings to be more intelligent through real-time monitoring and control. Today's BASs are becoming increasingly complex. A typical BAS usually stores hundreds of sensor measurements and control signals at each time step, which produces massive high dimensional data sets. Although massive data sets about the actual building operation are collected and stored in BASs, few attempts have been made to fully utilize such data. Tradition analysis methods for BAS data only focus on a small subset of the data, which resulting in a huge information loss. Currently, BASs can only perform some simple data analysis and visualization tasks, such as historical data tracking and moving average. The capability of systematically handling massive BAS data for more complex tasks is far from adequate.

Data mining (DM) is an emerging technology, which can effectively discover the hidden knowledge from large-scale data. Extensive explorations have been made to apply DM techniques in various fields, such as financial services, bioinformatics, retails, and telecommunication (Maimon and Rokach, 2010). However, DM applications in

the building field are less active. Previous research in the building field mainly adopted DM for three tasks, i.e., energy prediction (Dong *et al.*, 2005; Amin-Naseri and Soroush, 2008), fault detection and diagnosis (Yu *et al.*, 2012; Cabrera and Zareipour, 2013), and optimal controls (Kusiak *et al.*, 2010; Kusiak *et al.*, 2011). Although some encouraging results were obtained, the potential of DM in knowledge discovery of BAS data has not been fully explored. Previous work seldom took the large building operational data sets as a whole into consideration, and domain knowledge still plays the dominant role in the process of data analysis. Usually, only a small subset of BAS data was utilized in solving specific problems and those variables were generally selected based on domain expertise. For instance, the prediction model for chiller energy consumption may only adopts several input variables, e.g., part-load ratio, the supply and return temperature of chilled water, and the supply and return temperature of condenser water, as indicated by domain expertise. However, the significance of other variables, which may have a direct or indirect relationship to the building cooling load, cannot be discovered.

This study aims to analyze high dimensional BAS data using advanced data mining techniques, with the aim of identifying rare events in building operation. Rare event analysis is a highly efficient approach to gain insights from massive data sets, as it specifically focuses on the events which are very different from the majority. The causes of rare events in the building field may be unusual indoor and outdoor operating conditions and various faults occurring in building systems. Therefore, rare event analysis of the high dimensional BAS data helps to identify atypical building operating patterns, detect and diagnose faults, and eventually improve the building operational performance. Rare event analysis has been widely used for various applications, such as network intrusion detection, credit card fraud detection, and medical diagnostics (Lazerevic, et al., 2004).

In the building field, some attempts have been made on the use of rare event analysis to gain insights into building operation; however, some limitations do exist and further improvement is possible in three aspects. Firstly, the BAS data were not fully utilized. Previous research mainly used rare event analysis to find abnormality in building energy consumption (Seem, 2005; Seem, 2007; Liu *et al.*, 2010; Li, *et al.*, 2010; Khan *et al.*, 2013). Only a very small proportion of variables, which represent the system power consumption, were analyzed. In such a case, the rare events, which are only finable from a multivariate perspective, cannot be detected. Also, once a rare event is identified, a method should be developed to quickly identify which variables are causing such rarity. This is especially useful when the variable number is large. However, such effort is not seen in previous study. Secondly, previous research mainly adopts statistical methods, such as the generalized extreme studentized deviate, to detect rare events (Seem, 2007; Liu *et al.*, 2010; Li, *et al.*, 2010). The applicability of such methods is usually questionable when applied to high dimensional data, let alone whether the data meet the statistical assumptions or not. Rather than narrowing the analysis scope to a small amount of BAS variables, this research aims to develop a method to perform rare event analysis on high dimensional real-world BAS data. In such a case, the methods used should be scalable to high dimensional problems.

A holistic DM-based methodology is developed to specifically address the above-mentioned challenges. It consists of four steps, i.e., data preparation, rare event detection, rare event diagnosis, and post-mining. The first step uses subspace clustering analysis to identify typical building operating patterns. Then, outlier ensembles, which are formed by three advanced high dimensional outlier detection algorithms, are developed to detect rare events. Afterwards, a classification technique is used to output the variables that contribute more to the rarity. The last step interprets the results using domain knowledge and applies the findings to gain insights or improve building operational performance. The method is adopted to analyze the BAS data of the tallest building in Hong Kong, the International Commerce Center (ICC).

2. OVERVIEW OF DM TECHNIQUES

2.1 Research Outline

Figure 1 shows the schematic outline of this research. The proposed methodology consists of four stages. The first stages uses subspace clustering analysis to identify typical building operating patterns. One popular subspace clustering method, ORCLUS (Aggarwal and Yu, 2000), is employed to identity typical building operating patterns. Then, outlier detection ensembles are constructed using three advanced high dimensional outlier detection algorithms, i.e., angle-based outlier detection (ABOD) (Kriegel, et al., 2008), subspace outlier detection (SOD) (Kriegel, et al., 2009) and feature bagging-based method (FB) (Lazrevic and Kumar, 2005). These outlier detection ensembles are applied to find rare events in each cluster. The third stage applies random forests to identify variables contributing the most to the outlierness, or the rarity. The random forest algorithm (Breiman, 2001) is selected due to its excellent ability in handling high dimensional classification problems. The last stage uses domain knowledge to interpret results and thereby, using the discovered knowledge for building energy performance improvement.

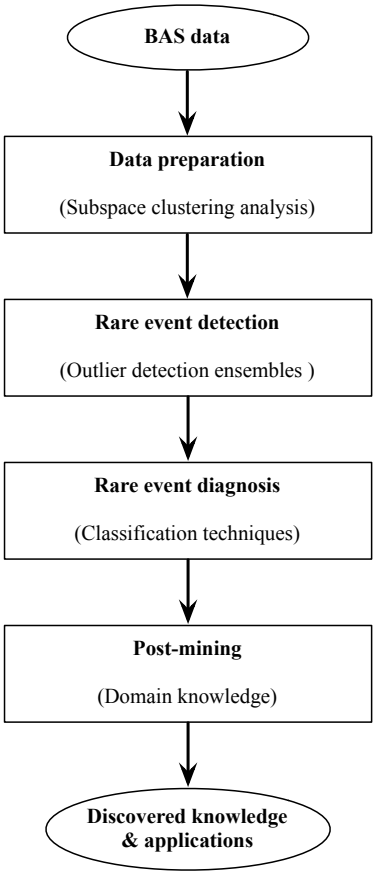

Figure 1: Schematic outline of the research

2.2 Subspace Clustering Analysis

Clustering analysis aims to discover groups or clusters, which contain objects with similar characteristics. The similarities between any pair of observations are normally evaluated using distance-based metrics, such as the Manhattan and Euclidean metrics. New challenges emerge as the dimension of data becomes larger and larger. The performance of clustering analysis in high dimensional space is negatively affected by two factors. Firstly, some variables may be irrelevant and therefore, the true cluster memberships may be masked if all dimensions are considered. Secondly, as data dimension increases, distance-based metrics become increasingly meaningless. Various solutions have been proposed to enhance the clustering efficiency in high dimensional space. In general, these methods can be divided into two categories (Parsons *et al.*, 2004), i.e., dimensionality reduction and subspace clustering. Dimensionality reduction methods can be further categorized into feature transformation and feature selection. Feature transformation aims to represent the data with fewer dimensions, but it is less effective as the relative distances are still preserved. Feature selection only selects the most relevant variables for clustering analysis. However, it is often difficult to make such selection without prior knowledge and such method does not perform well when clusters are to be found in different subspaces (Parsons *et al.*, 2004).

By contrast, subspace clustering integrates both feature evaluation and clustering to find clusters in different subspaces. It provides more flexibility in clustering as it enables users to breaking the assumption that all of the clusters are to be found in the same set of dimensions. It is especially useful in domains where one expects to find relationships across a variety of perspectives (Parsons *et al.*, 2004). In this study, one popular subspace clustering algorithm, i.e., ORCLUS, is employed. It can be summarized as a three-step approach. The first step iteratively assigns observations to the nearest cluster center. The second step redefines the subspace associated with each cluster by calculating the covariance matrix and selecting the orthonormal eigenvectors with the least spread (i.e., the smallest eigenvalue). The third step merges clusters that are near to each other and have similar directions of least spread. Two parameters are defined prior to algorithm execution, i.e., the number of clusters and the size of

subspace dimensionality. The choice of these parameters can be optimized using the cluster sparsity coefficient. More details can be found in (Aggarwal and Yu, 2000; Parsons *et al.*, 2004).

2.3 Outlier Detection Ensembles

Outliers, or rare events, are observations which deviates significantly from the others in the same data set. Outlier detection has been used in network intrusion detection, credit card fraud detection, and rare disease detection. In general, the outlier detection in massive BAS data faces two main challenges, i.e., the robustness and reliability of detection results, and the effectiveness in high dimensional space. In this study, these two challenges are addressed by developing outlier detection ensembles consisting of three advanced high dimensional outlier detection algorithms. Outlier detection ensembles integrate outlier detection and ensemble learning to enhance the reliability and robustness of outlier detection (Zimek et al., 2013).

ABOD uses the variance of angles between one observation and all the other pairs of observations to evaluate outlierness. It is effective since the angle is a more robust measure than the distance in high dimensional space (Kriegel *et al.*, 2008). If the variance of angles of an observation is large, such observation is more likely to be surrounded by others and therefore, it is less likely to be an outlier. By contrast, a small variance of angles indicates that most of the other observations are located in similar directions and therefore, such observation is more likely to be an outlier.

SOD (Kriegel *et al.*, 2009) uses the concept of axis-parallel subspaces to enhance the outlier detection performance in high dimensional space. Three parameters should be defined prior to the algorithm execution, i.e., k, which is the number of nearest neighbors to be considered when calculating the shared nearest neighbor similarity (it is measured based on the number of common nearest neighbors); l, which is the size of reference sets; and α, the coefficient used to calculate the threshold variance. The outlierness is evaluated by calculating the distance between the considered observation and the constructed subspace hyperplane.

FB (Lazarevic and Kumar, 2005) integrates the concept of ensemble learning to improve the outlier detection performance in high dimensional data sets. It can be summarized as a four-step method. Firstly, the size of feature subset is randomly selected from a uniform distribution between $d/2$ and $d-1$, where d is the data dimension. Then, the features are randomly selected without replacement. Thirdly, the classical density-based outlier detection method, i.e., local outlier factor (LOF) (Breunig *et al.*, 2000), is applied considering the feature subset. Such procedures are repeated N times and the final results are obtained through certain combination schemes.

Outlier scores generated by different outlier detection algorithms differ in their meaning, range, and contrast (Kriegel *et al.*, 2011). For instance, the outlier scores generated by ABOD range from 0 to infinity, and the smaller scores indicate a higher outlierness. By contrast, the classical density-based method, local outlier factor (LOF), outputs scores from 1 to infinity, and a larger score indicate a higher outlierness. The outlier score unification scheme proposed by Kreigel *et al.* (2011) is adopted in this study. It consists of two steps, i.e., regularization and normalization. Regularization aims to transform the outlier scores onto the interval between 0 and infinity, and a larger regularized score indicate a higher outlierness. Normalization aims to transform the regularized scores onto the interval between 0 and 1, and such normalized scores can be interpreted as the probability of being an outlier. In this study, ABOD scores are regularized through logarithmic inversion, i.e., $-log(S_i/S_{max})$, where S_i is the outlier score of i^{th} observation, and S_{max} is the maximum outlier score returned by ABOD. FB scores are regularized by subtracting 1 from the original FB scores. No regularization is needed for SOD scores. All the regularized scores are transformed to outlier probability by the Gaussian scaling, i.e., $max\{0, erf((S_i-\mu_s)/(\eta\sigma_s))\}$, where $erf()$ stands for the Gaussian error function, S_i is the outlier score for the i^{th} observation, μ_s is the mean of outlier scores, η is a coefficient for Gaussian scaling, and σ_s is the standard deviation of outlier scores.

2.4 Random Forests

Random forest is developed by Breiman (2001) to handle both classification and regression problems. In essence, random forests are ensembles and their base models are generated using two randomization strategies, i.e., each tree is trained using a random set of observations, and a random subset of features is used for each tree node split. It has been shown to have excellent ability in handling high dimensional problems (Breiman, 2001).

One important by-product of random forests is the variable importance. It has been widely used for variable selection. In general, there are three types of approaches to variable importance evaluation (Strobl *et al.*, 2007). The first type is rather native, which uses the number of times each variable being selected by all individual trees as the indicator. The second type incorporates a weighted mean of the individual trees' improvement in the splitting criterion generated by each variable (Friedman, 2001). The third type is called the permutation accuracy importance.

The rationale behind is that random permutation of a predictor X_i can break its original association with the response Y. If X_i is associated with Y originally, the accuracy after permutation is expected to decrease. Thus, the difference in accuracy before and after permutation can be used to indicate variable importance. It is argued that the first and second types usually result in biased outcomes, favoring continuous variables and variables with many categories. By contrast, the permutation-based method is able to provide unbiased results (Friedman, 2001). In this study, the random forest algorithm is used as the classification tool, and the permutation-based method is used to output unbiased variable importance.

3. RARE EVENT ANALYSIS OF HIGH DIMENSIONAL BAS DATA

3.1 Description of the building system and raw BAS data

The data used in this study were collected from the highest commercial building in Hong Kong, i.e., International Commerce Center (ICC). An advanced BAS is installed in this building. Over 1200 sensor measurements or control signals are stored in the BAS data with a collection interval of 1 minute. This study focuses on the central chilling system of the heating, ventilation, and air-conditioning (HVAC) system. The central chiller system is briefly introduced as follows. The building is divided into 4 zones to prevent the chilled water pipelines and terminal units from suffering extremely high pressure (i.e., the highest static pressure of more than 40 bar and the designed working pressure of nearly 60 bar). 6 water-cooled chillers, each has a cooling capacity of 7230 kW and a rated power consumption of 1346 kW, are installed to provide cooling source. Each chiller is associated with one constant condenser water pump and one constant primary chilled water pump. 11 evaporative cooling towers with a total design capacity of 51,709kW are used to reject heat from chiller condensers. The cooling energy to the demand side is distributed by the use of 12 variable-speed secondary chilled water pump (SCHWP) and heat exchangers. The chilled water for zone 2 is supplied by secondary chilled water pumps directly. Heat exchangers are used to distribute cooling energy for the other zones.

In total, 3-month data (from April, 2013 to June, 2013) were used in this study, resulting in 123,120 observations in total. To reduce the computation load, the raw data is aggregated to a 15-minute time interval by taking the mean. Observations with missing values are neglected. The resulting data set consists of 7752 observations and 187 variables, e.g., the date and time (i.e., Year, Month, Day, Hour, Minute, Weekday), indoor and outdoor conditions (e.g., outdoor temperature and relative humidity), power consumption of different components (e.g., water-cooled chillers, cooling towers, chiller water pumps), flow rate, pressure, and temperature at different key positions of HVAC central chilling system.

3.2 Identification of typical building operating patterns

The HVAC system is highly dynamic and its operating variable varies dramatically under different building operating patterns. The identification of typical operating conditions is a must for rare event detection, as it provides reference sets to correctly evaluate the rarity. This study uses the power consumptions of the key components in the central chilling system to identify typical building operating patterns. As introduced in section 3.1, the operating patterns of the ICC central chilling system can be revealed by analyzing the power consumptions of three main components, i.e., chillers, cooling towers, and secondary chilled water pumps. It is noted that neither the primary chilled water pumps nor the condenser water pumps are considered. These two kinds of pumps are both constant-speed pumps and their power consumptions quickly become constant once the associated chillers are switched on. Therefore the information conveyed by these variables is redundant. To summarize, the data prepared for clustering analysis have 26 variables, i.e., the power consumptions of 6 chillers, 11 cooling towers, and 9 secondary chilled water pumps (the quality of the measurements for SCHWPs No.1, 2 and 12 are very poor and hence are excluded).

Prior to the execution of clustering algorithm, normalization is performed to prevent the bias towards variables with larger scales. Max-min normalization, i.e., $(x_i-x_{min})/(x_{max}-x_{min})$, is used to scale the data to a range between 0 and 1. The subspace clustering algorithm ORCLUS is then applied. Two parameters, i.e., the cluster number and the size of subspace dimensionality, are optimized based on the sparsity coefficient, and are 9 and 5, respectively. The numbers of observations in each cluster and the corresponding percentages are shown in Table 1.

As a summary of the observations in each cluster, the aggregated power consumptions (kW) of each sub-system and the standard deviation are calculated and reported in Table 2. The results show that subspace clustering is effective in revealing typical operating patterns. Compared to conventional methods which calculate the distance based on full dimensional space, subspace clustering is especially useful when clusters are expected to be found across different perspectives. One may notice for Clusters 4 and 7, both the mean aggregated power consumptions of three sub-systems and their corresponding standard deviations are quite similar. In such a case, conventional methods tend

to merge these two clusters as one. However, the subspace clustering method successfully identifies the subtle differences, i.e., even though the aggregated power consumption levels of individual components are similar, the operating units are slightly different. For instance, CTs 2, 4, 5, 6, 7, 9,10 and SCHWPs 3, 7, 9 tend to be under operation for observations in Cluster 4. While for Cluster 7, the operating units tend to be 1, 2, 5, 6, 7, 8, 10 and SCHWPs 3, 8, 10. The ability of distinguishing such subtle differences is essential in preparing more suitable reference sets for rare event detection.

Table 1: Cluster information

Cluster	1	2	3	4	5	6	7	8	9
Sizes	634	969	1599	1058	454	1104	693	704	537
Percentage (%)	8.18	12.50	20.63	13.65	5.86	14.24	8.94	9.08	6.93

Table 2: Summary of clustering results

Cluster	WCC Mean	WCC SD	CT Mean	CT SD	SCHWP Mean	SCHWP SD
1	1319.75	1000.81	542.55	258.97	132.60	57.38
2	1622.72	948.50	313.62	183.82	117.37	59.84
3	1128.37	968.34	402.04	240.56	114.70	54.48
4	1508.30	740.19	256.49	88.98	85.83	21.36
5	1703.37	1094.35	298.74	173.32	84.35	45.55
6	1220.15	505.76	212.19	85.75	72.07	22.16
7	1597.65	787.43	295.00	100.62	89.38	27.12
8	2204.30	942.76	405.85	221.42	128.97	64.61
9	2525.80	670.80	812.59	56.13	171.72	36.16

3.3 Rare event detection

The developed outlier detection ensemble has 31 base detectors, i.e., 1 ABOD detector, 25 SOD detectors, and 5 FB detectors. SOD detectors are developed using different parameter settings of k, l, and α. Both k and l have 5 possible values, i.e., 5, 10, 15, 20, 25. α is fixed as 0.8, as recommended by Kriegel *et al.* (2009). FB detectors are developed using different q, i.e., 5, 10, 15, 20, 25. The number of iterations for FB is fixed as 500. The raw scores are transformed using the outlier score unification scheme described in section 2.3. The final outlier probability is obtained by calculating the weighted averaged probability.

Table 3: Summary of rare event percentages in each cluster

Cluster	1	2	3	4	5	6	7	8	9
Percentage	1.71%	1.76%	1.58%	1.99%	1.85%	1.82%	2.05%	2.63%	1.46%

Rare events are detected by comparing the final outlier probability and the probability threshold, i.e., 0.75. Table 3 summarizes the percentage of observations in each cluster having a probability larger than the threshold. Results show that detected rare events only account for a small proportion, ranging from 1.46% (Cluster 9) to 2.63% (Cluster 8). Figure 3 is presented as an example to show the relative frequency distributions of final outlier probabilities in Cluster 8. The vertical dotted line indicates the specified threshold, 0.75. In general, the relative frequency drops as the probability score increases. It shows that nearly 30% of the observations in Cluster 8 are definitely not rare events, as the outlier probability is 0. Around 55% of the observations are very unlikely to be outliers as the resulting probabilities are less than 0.35. The outlier probabilities of other 15% observations are uniformly distributed between 0.35 and 0.95.

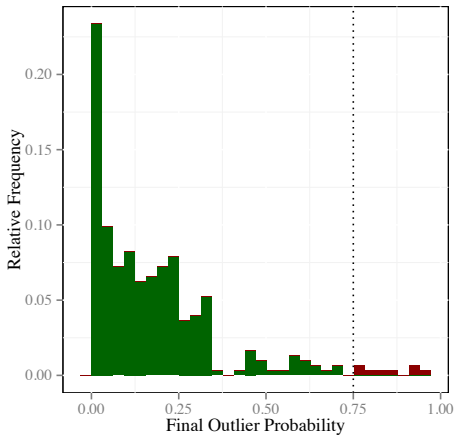

Figure 3: Relative frequency distribution of final outlier probabilities in Cluster 8

3.4 Rare event diagnosis

Rare event diagnosis aims to effectively identify variables causing the rarity. The random forest algorithm is employed to classify rare events against normal ones. The resulting variable importance indicates the contribution to the rarity. Firstly, for each detected rare events, a reference set is constructed by selecting *m* normal observations, which have the top *m* shared nearest neighbor similarities to the considered rare event. The value of *m* is chosen as 25. Then, a simple oversampling method, which replicates the considered rare event *(m-1)* times, is used to make it a class-balanced classification problem. Thirdly, random forest is applied to classify the data. One parameter, which specifies the number of variables to be considered at each split, is optimized through 10-fold cross validation. The Cohen's Kappa coefficient is applied to evaluate the classification accuracy. Such metric is more robust than the simple percent agreement as it takes into account the agreement occurring by chance (Maimon and Rokach, 2010).

4. POST-MINING

4.1 Identification of system transient operation

The observation recorded at 20:30 on June 7, 2013 belongs to the 9th cluster and is identified as a rare event with an outlier probability of 0.79. The implementation of random forest algorithm classifies such observation against the reference set, with a Kappa accuracy of 0.94. The variables resulting in the top 5 variable importance are the condenser water flow rate of Chiller 4 (denoted as Flow_COND_WCC_04), the chilled water flow rate of Chiller 4 (denoted as Flow_EVAP_WCC_04), condenser water flow rate of Cooling Towers 2, 3, and 10 (denoted as Flow_CT_02, 03, and 10 respectively).

Table 4 compares the means of these 5 variables in reference set and the values of the rare event. It is obvious that significant differences exist between the mean values in the reference set and the values of the rare event. More specifically, the mean values of the condenser water and the chilled water flow rate indicate that Chiller 4 is running at a near full-load condition for observations in the reference set. It is consistent with the results obtained in clustering analysis, as the power consumption of Chiller 4 in Cluster 9 is very close to the rated power consumption. However, the values of these two variables drop to around zero in the rare event observation. Similarly, the other three variables show that Cooling Tower 2, 3, and 10 tend to run at a moderate level in the reference set, while dramatic decreases are observed in the rare event observation. Further check ensures that there is little difference in the flow rates of the other 5 chillers between the reference set and rare event. Therefore, it is reasonable to claim that the central chilling system is undergoing a stage-down process, and in this specific case, Chiller 4 is being switched off. Such kind of rare events actually represents system transient operations.

It is observed that around 85% of the detected rare events are related to system transient operations. By analyzing such rare events, insights into the building operating behaviors can be gained. For instance, the diagnosis results show that two rounds of stage-up and stage-down are normally used for daily operation, each lasts from 30 minutes to 1 hour. The stage-up process normally takes place at 4 a.m. to 5 a.m. and 7 a.m. to 8 a.m., while the stage-down process takes place at 8 p.m. to 9 p.m. and 11 p.m. to 12 a.m. An additional stage-down process is observed between 1:30 p.m. and 2:30 p.m. for Saturdays. Such results are in accordance with domain knowledge, as many offices in Hong Kong do work half-day on Saturdays.

Table 4: Value comparison of variables contributing to the rarity

Comparison	Flow_COND_ WCC_04 (l/s)	Flow_EVAP_ WCC_04 (l/s)	Flow_CT_ 02 (l/s)	Flow_CT_ 03 (l/s)	Flow_CT_ 10 (l/s)
Reference	333.97	357.00	115.12	65.82	88.07
Rare event	2.00	0.67	75.97	41.60	54.50

4.2 Identification of inappropriate sensor installation

The observation recorded at 14:15 on June 8, 2013 belongs to the 2nd cluster and is identified as a rare event with an outlier probability of 0.76. The variable importance retuned by the random forest model shows that the Chiller 5's chilled water flow rate (Flow_EVAP_WCC_05), the condenser water flow rate (Flow_COND_WCC_05), the pressure difference on the condenser side (PresD_COND_WCC_05), and the pressure difference on the evaporator side (PresD_EVAP_WCC_05) are the most significant variables causing the rarity.

Table 5: Value comparison of variables contributing to the rarity

Comparison	Flow_COND_ WCC_05 (l/s)	Flow_EVAP_ WCC_05 (l/s)	PresD_COND_ WCC_05 (kPa)	PresD_EVAP_ WCC_05 (kPa)
Reference	317.73	358.15	114.84	118.32
Rare event	4.80	2.81	315.56	8.24

Table 5 shows the mean values of these variables in the reference set and the values of the rare event. It is apparent that the rare event stands for the status when Chiller 5 is nearly switched off, as both its condenser flow rate and the chilled water flow rates are close to 0. The reading of the pressure difference on the evaporator side is in accordance with such change, as it also approaches 0. However, the pressure difference across the condenser side is much larger than that when Chiller 5 is in full operation. Such observation clearly violates the domain knowledge.

There are two possible explanations for this phenomenon. The first is sensor fault. If this is the case, the sensor may have lost its functionality, and the resulting abnormal values are expected to last for a relatively long time until maintenance (e.g., sensor replacement or repair) are introduced. Such explanation is not applied to this case, as the value of PresD_COND_WCC_05 soon becomes normal when Chiller 5 is brought back into operation in one and a half hour. Further investigation shows that PresD_COND_WCC_05 is always around 300 kPa when Chiller 5 is idle and is always around 120 kPa when Chiller 5 is in operation. It indicates that the sensor is functioning properly.

The second explanation is suboptimal sensor installation. Figure 4 is depicted to illustrate this issue. The pressure difference across the condenser is obtained by taking the difference in readings between two sensors, which are allocated at either side of the condenser. Conventionally, these two sensors are installed at point B and C. In such a case, the measured pressure difference is normal whenever the chiller is switched on or switched off. The observed situation indicates that these two sensors are not installed properly. Instead of being installed at point B, the sensor measuring the incoming condenser water flow is installed at point A, which is before the control valve. As shown in Figure 2, the condenser water system of six chillers is connected in parallel. Consequently, when Chiller 5 and its associated condenser water pump are switched off, the pressure at point A is actually very similar to the pressure at the same position of the other branches. Meanwhile, the pressure at point C is around 0 since both valves have been closed. In such a case, the measured pressure difference is very large when chiller is switched off and normal when chiller is switched on. Hence, the observed rare event indicates that the sensor installation is suboptimal for Chiller 5 and further actions are needed to avoid confusion.

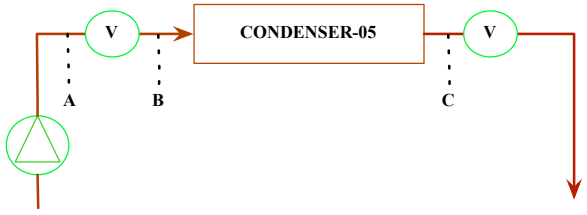

Figure 4: Suboptimal sensor installation

5. CONCLUSIONS

The advance in technology enables a more comprehensive monitoring and control system to regulate building operational performance. The resulting BAS data are normally of great complexity and massive volume. Advanced approaches, which integrate both domain expertise and modern analytics, are desired to make full use of BAS data. Rare event analysis on BAS data is an efficient approach to extract useful knowledge, as it focuses on particular events of interest. It helps to evaluate building operational performance, identify potential faults in design and operation, and understand building operating behaviors.

This paper presents an applicable approach to rare event analysis for BAS data, with the focus of enhancing the effectiveness when the data dimension is high and little prior knowledge is available. To correctly evaluate the rarity of each observation, a suitable reference set should be selected for comparison. Therefore, the first step of this approach focuses on the identification of typical building operating patterns. Among various clustering techniques, the subspace clustering algorithm is selected as it is expected to find clusters across different subspaces. The research results validate such choice, as it successfully distinguishes the subtle difference when the power consumption levels are similar but the running units are different. The second step detects rare events without explicitly specifying the candidate variables. In another word, rather than focusing on a small subset of variables that are selected by domain expertise or experience, this approach aims to examine the rarity using the whole available data. Therefore, the data is fully utilized and unexpected knowledge can be revealed. The adoption of the whole data set imposes a question on the efficiency and effectiveness of conventional rare event detection methods. In this study, three advanced outlier detection methods, which are specially designed for high dimensional problems, are selected as the detection algorithms. Ensemble learning is integrated to further enhance the robustness and reliability of detection results. The proposed approach also includes a diagnosis step. The classification technique is used to classify the rare events against the normal observations, and the resulting variable importance is extracted as indicators. The random forests method is selected, as it is suitable of providing reliable results for high dimensional problems.

The results show that useful knowledge on system transient changes, building operating behaviors, and potential faults in design and operation can be obtained. More specifically, the majority of the identified rare events are observations recorded when the system is undergoing transient changes, e.g., ON/OFF of chillers and cooling towers. Such obtained knowledge helps to understand the building operating behaviors (e.g., stage-up and stage-down schedules for daily operation), and interactions between individual components or subsystems. Observations with abnormal readings are also successfully identified. Domain knowledge is involved to decide whether the abnormality is due to sensor faults, operating faults, or design/installation faults. Accordingly, further actions can be taken to improve the building operational performance. Moreover, the obtained knowledge can be used to as the basis for other studies. For instance, many studies, e.g., component-level or system-level performance assessments, rely on the collection of steady-state data. In such cases, the inclusion of any observations under system transient changes may greatly distort assessment results. The proposed approach helps to automatically identify such observations with little prior knowledge and its significance emerges as the BAS data becomes more complex and high dimensional.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support of this research by the Hong Kong Polytechnic University (project No. G-YM86).

REFERENCES

Aggarwal, C.C., Yu, P.S., 2000, Finding generalized projected clusters in high dimensional spaces, *In Proceedings of the ACM SIGMOD International Conference on Management of Data 2000*, Dallas, Texas, USA.

Amin-Naseri, M.R., Soroush, A.R., 2008, Combined use of unsupervised and supervised learning for daily peak load forecasting, *Energ. Convers. Manage.*, vol. 49, no.6: p. 1302-1308.

Breiman, L., 2001, Random forests, *Mach. Learn.*, vol. 45, no. 1: p. 5-32.

Breunig, M.M., Kreigel, H.P., Ng, R., Sander, J., 2000, LOF: Identifying density-based local outliers, *In Proceedings of the ACM SIGMOD International Conference on Management of Data 2000*, Dallas, Texas, USA.

Cabrera, D.F.M., Zareipour, H., 2013, Data association rule mining for identifying lighting energy waste patterns in educational institutes, *Energ. Buildings.*, vol. 62, no. 26: p. 210-216.

Dong, B., Cao, C., Lee, S.E., 2005, Applying support vector machines to predict building energy consumption in tropical region, *Energ. Buildings.*, vol. 37, no. 5: p. 545-553.

Friedman, J., 2001, Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, vol. 29, no. 5: p. 1189-1232.

Hong Kong energy end-use data 2013, Hong Kong Electrical & Mechanical Services Department (EMSD), September 2013.

International Energy Agency (IEA), http://www.iea.org/aboutus/faqs/energyefficiency/, accessed on Jan 22, 2014.

Khan. I., Capozzoli, A., Corgnati, S.P., Cerquitelli, T., 2013, Fault detection analysis of building energy consumption using data mining techniques, *Energ. Procedia.*, vol. 42, no. 57: p. 557-566.

Kriegel, H.P., Kroger, P., Schubert, E., Zimek, A., 2011, Interpreting and unifying outlier scores, *In Proceedings of the 11th SIAM International Conference on Data Mining*, Mesa, Arizona, USA.

Kriegel, H.P., Kroger, P., Schubert, E., Zimek, A., 2009, Outlier detection in axis-parallel subspaces of high dimensional data, *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, Bangkok, Thailand: p. 831-838.

Kriegel, H.P., Schubert, M., Zimek, A., 2008, Angle-based outlier detection in high-dimensional data, *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Las Vegas, Nevada, USA: p. 444-452.

Kusiak, A., Li, M.Y., Tang, F., 2010, Modeling and optimization of HVAC energy consumption, *Appl. Energ.*, vol. 87, no. 10: p. 3092-3102.

Kusiak, A., Tang, F., Xu, G.L., 2011, Multi-objective optimization of HVAC system with an evolutionary computation algorithm, *Energ.*, vol. 36, no. 5: p. 2440-2449.

Lazarevic, A., Kumar, V., 2005, Feature bagging for outlier detection, *In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Chicago, Illinois, USA: p. 444-452.

Lazarevic, A., Srivastava, J., Kumar, V., 2004, 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*.

Li, X., Bowers, C.P., Schnier, T., 2010, Classification of energy consumption in buildings with outlier detection, *IEEE Transactions*, vol. 57, no. 11: p. 3639-3644.

Liu, D., Chen, Q., Mori, K., Kida, Y., 2010, A method for detecting abnormal electricity energy consumption in buildings, *J. Comput. Info. Sys.*, vol. 6, no. 14: p. 4887-4895.

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

Parsons, L., Haque, E., Liu, H., 2004, Subspace clustering for high dimensional data: A review, *ACM SIGKDD Exploration Newsletter-Special Issue on Leaning from Imbalanced Datasets*, vol. 6, no. 1: p. 90-105.

Seem, J.E., 2005, Pattern recognition algorithm for determining days of the week with similar energy consumption profiles, *Energ. Buildings.*, vol. 37, no. 2: p. 127-139.

Seem, J.E., 2007, Using intelligent data analysis to detect abnormal energy consumption in buildings, *Energ. Buildings.*, vol. 39, no. 1: p. 52-58.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007, Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC. Bioinformatics.*, vol. 8, no.1: p. 8-25.

Yu, Z., Haghighat, F., Fung, C.M., Zhou, L., 2012, A novel methodology for knowledge discovery through mining associations between building operational data, *Energ. Buildings.*, vol. 47, no. 50: p. 430-440.

Zimek, A., Gaudet, M., Campello, R.J.B., Sander, J., 2013, Subsampling for efficient and effective unsupervised outlier detection ensembles, *In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Chicago, Illinois, USA.