Applied Energy 127 (2014) 1–10

Contents lists available at ScienceDirect

Applied Energy

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

Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques

Cheng Fan, Fu Xiao *, Shengwei Wang

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

HIGHLIGHTS

- A data mining based method is proposed to predict building energy consumption.
- The outlier detection method can identify abnormal building operating patterns.
- He recursive feature elimination technique is effective in selecting optimal inputs.
- The prediction performances of eight popular predictive algorithms are studied.
- Ensemble models built on the eight base models have the best performances.

ARTICLE INFO

Article history:
Received 28 November 2013
Received in revised form 5 March 2014
Accepted 6 April 2014
Available online 26 April 2014

Keywords:
Building energy prediction
Data mining
Feature extraction
Clustering analysis
Recursive feature elimination
Ensemble model

ABSTRACT

This paper presents a data mining (DM) based approach to developing ensemble models for predicting next-day energy consumption and peak power demand, with the aim of improving the prediction accuracy. This approach mainly consists of three steps. Firstly, outlier detection, which merges feature extraction, clustering analysis, and the generalized extreme studentized deviate (GESD), is performed to remove the abnormal daily energy consumption profiles. Secondly, the recursive feature elimination (RFE), an embedded variable selection method, is applied to select the optimal inputs to the base prediction models developed separately using eight popular predictive algorithms. The parameters of each model are then obtained through leave-group-out cross validation (LGOCV). Finally, the ensemble model is developed and the weights of the eight predictive models are optimized using genetic algorithm (GA).

The approach is adopted to analyze the large energy consumption data of the tallest building in Hong Kong. The prediction accuracies of the ensemble models measured by mean absolute percentage error (MAPE) are 2.32% and 2.85% for the next-day energy consumption and peak power demand respectively, which are evidently higher than those of individual base models. The results also show that the outlier detection method is effective in identifying the abnormal daily energy consumption profiles. The RFE process can significantly reduce the computation load while enhancing the model performance. The ensemble models are valuable for developing strategies of fault detection and diagnosis, operation optimization and interactions between buildings and smart grid.

1. Introduction

The ever-increasing energy consumption has raised worldwide concern over the issues of environmental degradation, energy security and geopolitics. According to statistics from the International Energy Agency (IEA), buildings are responsible for 32% of total final energy consumption [1]. In terms of primary energy consumption, buildings represent around 40% in most IEA countries [1]. Such figures would be much higher in less industrial-oriented districts. For example, the building sector in Hong Kong accounts for over 60% of the total final energy consumption and over 90% of electricity use [2]. Building energy efficiency is of great importance to global sustainability.

In the past two decades, researchers have devoted themselves in the improvement of building energy efficiency. Building energy consumption prediction has drawn special attention as it is often needed in developing various strategies for improving building energy performance, e.g., fault detection and diagnosis [3], demand side management for smart grid [4]. According to the time-scale of prediction, the existing research on building energy prediction can be broadly classified into three categories, i.e., short-term (i.e., up

* Corresponding author. Tel.: +852 2766 4194; fax: +852 2765 7198.
E-mail address: linda.xiao@polyu.edu.hk (F. Xiao).

http://dx.doi.org/10.1016/j.apenergy.2014.04.016
0306-2619/

to one week ahead), medium-term (i.e., from one week to one year ahead) and long-term (i.e., longer than one year ahead) predictions [5]. Currently, the short-term prediction is the main focus due to its close linkage to the day-to-day operations [6]. The short-term prediction mainly focuses on the predictions of the daily peak demand, daily energy consumption and daily load profiles [7]. Popular methods for developing prediction models include the engineering methods [8,9], statistical methods [10,11], gray-box modeling methods [12,13], machine learning and artificial intelligence methods [14–18].

Nowadays, buildings are becoming not only energy-intensive, but also information-intensive. The rich data provides convenience in modeling complex and nonlinear processes in building operations. Meanwhile, it can be tedious and time-consuming to process large-scale data. In terms of short-term building energy predictions, the existing research needs improvement in three aspects. Firstly, before the development of prediction models, efficient and effective methods should be developed to enhance the quality of massive building energy consumption data. Secondly, existing research mainly utilized domain knowledge or conventional filter methods to select input variables. Consequently, some useful knowledge may be overlooked and the resulting models may not work well under different conditions. To overcome such shortcomings, a data-driven input selection process is proposed in this study. Thirdly, individual predictive algorithms have their own pros and cons. A more advanced data mining technique, ensemble learning, can develop composite models to improve the accuracy and stability of predictions.

This paper develops an effective approach to constructing the prediction models of the next-day energy consumption and peak power demand, taking into account the three deficiencies of existing research mentioned above. Abnormal building energy consumption profiles are firstly identified and removed using feature extraction, clustering analysis, and the generalized extreme studentized deviate (GESD). Base models are then developed using eight popular predictive algorithms. A data-driven input selection algorithm, the recursive feature elimination (RFE), is applied to find inputs to the eight base models separately. The ensemble models are constructed by combining eight base models. Genetic algorithm (GA) is used to optimize the weights of eight base models in the final ensembles. The proposed approach is applied to analyze the large energy consumption data of the tallest building in Hong Kong. The performances of individual base models and the ensemble models, as well as their computation times are compared.

2. Description of data mining techniques

2.1. Research outline

Fig. 1 shows the schematic outline of the research. One year building energy consumption data, collected at 15-min intervals, are adopted for analysis. The data preparation contains three main tasks, i.e., data transformation, feature extraction, and creation of candidate input pool. The identification of abnormal building energy consumption profiles is achieved by using clustering analysis and outlier detection. The entropy-weighted *k*-means (EWKM) algorithm is used as the clustering algorithm and the GESD algorithm is adopted for outlier detection. The RFE is performed to select the optimal input variables for different predictive algorithms. Model parameters are optimized through leave-group-out cross validation (LGOCV). Then, GA is used to optimize the weights of eight base models in ensemble models, which output the final prediction results of the next-day daily peak power demand and daily energy consumption.

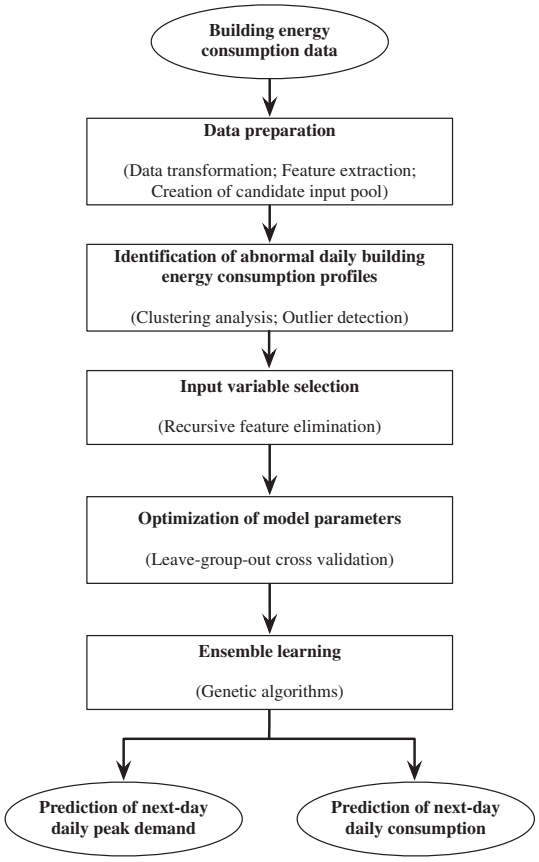

Fig. 1. Schematic outline of the research.

2.2. Clustering analysis

Clustering analysis aims to group observations with similar characteristics within the same cluster. The similarities between any pairs of observations are normally evaluated using distance-based metrics, such as the Manhattan and Euclidean metrics. The desired clustering result aims to maximize observation similarities within the same cluster and minimize the similarities between different clusters. Clustering analysis has been successfully used to preprocess large data sets, identify outliers and discover underlying patterns [20]. In this study, the entropy weighted *k*-means (EWKM) method is adopted to identify typical building energy consumption profiles. Three parameters should be specified to perform the EWKM algorithm, i.e., the cluster number (k), the weight distribution parameter (λ), and the convergence threshold (δ). The optimal parameter values can be determined using either internal validation methods (e.g., Davies–Bouldin index, Silhouette index and Dunn index) or external validation methods (e.g., Purity, *F*-measure and Normalized mutual information) [20]. The details of EWKM algorithm can be found in [19].

2.3. Generalized extreme studentized deviate (GESD)

Outliers are observations which appear to be inconsistent with the remainder of a specific data set [21]. Outliers may arise due to various reasons, such as human error, instrument error and change of system behavior. Among the existing algorithms, the GESD algorithm [22] was highly recommended because of its flexibility under various conditions [23]. It has been implemented in detecting abnormality in building energy consumption data and proved to be computationally efficient in handling large building energy data [3,24].

In this study, the GESD algorithm is employed to detect outliers in feature space. The basic steps of GESD can be briefly described as follows. Prior to the execution, users should define two values. One is the probability of Type I error, α, which determines the probability of incorrectly declaring a normal observation as an outlier. The other is the upper limit of potential outlier number, denoted as N. Firstly, the mean value and the standard deviation of the data set are computed. Then, N extreme observations, which result in the top N furthest deviations from the mean, are selected. For each extreme value X_i, where i ranges from 1 to N, the extreme studentized deviate Y_i is computed using Eq. (1). Meanwhile, the critical value for the ith extreme value is computed using Eqs. (2) and (3). Finally, Y_i is compared with the critical value. If Y_i is the larger one, then X_i is an outlier. Otherwise, X_i is not.

$$Y_i = \frac{|X_i - \overline{X}|}{SD} \tag{1}$$

$$\lambda_i = \frac{(n-i) \times t_{n-i-1,p}}{\sqrt{(n-i+1) \times (n-i-1+t^2_{n-i-1,p})}} \tag{2}$$

$$p = \frac{\alpha}{2 \times (n-i+1)} \tag{3}$$

where n is the total number of observations; $t_{n-i-1,p}$ is the Student's t-distribution with a degree of freedom equals to $(n-i-1)$, p is the probability corresponding to the critical value.

2.4. Recursive feature elimination (RFE)

The selection of input variables to a prediction model is very important when the number of candidate inputs is large and the prediction algorithms are complex. It helps to minimize the risk of over-fitting, reduce the computation costs, retain or modestly improve the model performance, and identify the intrinsic dimensionality of a given problem [25].

In general, there are two commonly used approaches to selecting model inputs. The first one is based on feature reconstruction (e.g., principal component analysis [20]). By projecting onto the first few principal directions, a new set of data with lower dimensions is obtained through the linear combination of original data. One disadvantage of such method is that none of the original data can be abandoned and it may be difficult to interpret the inputs [25]. The second one relies on the concept of subset selection. Commonly used methods can be further classified into filter, wrapper, and embedded methods. The filter method ranks the variables according to certain univariate metrics, such as the Pearson correlation coefficient. The disadvantage lies in the redundancy of the subset selected. Wrapper method evaluates the usefulness of subsets by considering a certain learning algorithm. Since exhaustive searches of subsets are performed, wrapper method may have a dramatic increase in computation costs. Alternatively, embedded method, which also performs variable selection based on certain learning algorithm, is more efficient since it is carried out by directly optimizing a two-part objective function with a goodness-of-fit and a penalty for a large number of input variables [25].

This study adopts RFE, an embedded method, to select the inputs to predictive models. In essence, RFE uses the backward selection technique. Firstly, one should determine the learning algorithm. Then, a model is trained by taking all candidate variables into account. Certain ranking criterion, which evaluates the variable importance, is computed for all the variables. Both the variable ranking and the model performance are stored for the final variable selection. The variable which results in the smallest ranking is then removed. This subset of input variables is updated for the next cycle of training. The procedure is conducted iteratively until there is no further variable to be removed. A more detailed elaboration can be found in [25].

2.5. Predictive algorithms

In this section, a brief overview about the adopted predictive algorithms is given. In total, eight widely used predictive algorithms, i.e., multiple linear regression (MLR), autoregressive integrated moving average (ARIMA), support vector regression (SVR), random forests (RF), multi-layer perceptron (MLP), boosting tree (BT), multivariate adaptive regression splines (MARS), and k-nearest neighbors (kNN) methods are used to map the relationship between inputs and outputs. These algorithms are selected based on two main considerations, i.e., popularity and diversity. All the selected algorithms have been widely used in solving complex modeling and prediction problems, and their performances are reported to be encouraging. In addition, they are selected to maximize the ensemble diversity, which is benefit to the ensemble performance [30]. Each selected predictive algorithm has its own advantages and specialties. For instance, the ARIMA method is effective in mapping linear relationship, while SVR is well known for its ability in capturing nonlinearity.

The first two linear methods, i.e., the multiple linear regression (MLR) and autoregressive integrated moving average (ARIMA), are selected as performance benchmarks. MLR is one of the most widely used methods to perform linear regression analysis. ARIMA is the most general technique for predicting time series data. An ARIMA model is normally defined by three parameters, i.e., the number of autoregressive terms p, the number of difference d, and the number of moving average terms q. More detailed explanations can be found in [26].

Support vector regression (SVR) is developed by Vapnik in 1995 to handle regression problems. SVR uses kernel function to solve nonlinear problems more efficiently. In this study, the Gaussian radial basis function is adopted as the kernel function, as it is very effective and efficient in handling nonlinear problem [16]. A more comprehensive discussion on SVR is given in [20].

Random forest (RF) is developed by Breiman in 2001 for both classification and regression problems [27]. Two randomization strategies are normally used. First, each tree is trained by considering a random subset of observations. Second, a random subset of variables is considered to split each tree node. RF has been shown to be especially useful in handling high dimensional problem and the generalization performance is very competitive [20,27].

Multi-layer perceptron (MLP) is a feed-forward artificial neural network which is capable of solving both linear and nonlinear problems. It normally consists of a number of nodes organized in several layers. The weights of each node are adaptively adjusted using the back-propagation technique. MLP has been widely used in pattern recognition, prediction, and function approximation [20].

Boosting tree (BT) integrates the use of boosting methods with regression trees to solve classification or prediction problems. A sequence of simple tree models are developed in such a manner that each successive tree aims to model the residuals of the preceding tree [20]. The final model can be regarded as a weighted additive binary tree model [28].

Multivariate adaptive regression splines (MARS) is an adaptive non-parametric regression method [29]. MARS normally divides the input space into several sub-regions and develops separate models accordingly. More detailed discussions can be found in [20,29].

k-Nearest neighbors (kNN) is a non-parametric learning algorithm used for either classification or prediction. It is non-parametric as it does not learn an explicit mapping relationship between inputs and outputs. The final scores of testing data are generated using the proximity of neighboring inputs in the training data

and their corresponding outputs. The parameter, k, which defines the number of considered neighboring observations, should be determined prior to the scoring. kNN is regarded as one of the simplest learning algorithm, yet the performance in practical applications can be satisfactory [20].

2.6. Ensemble learning

Ensemble learning is a powerful machine learning method which integrates a number of base models to generate the final output. It has gain great popularity due to its excellent generalization performance. Ensemble models often result in much better performance than the individuals that make them up. The reasons behind are threefold [30]. Firstly, the training data may not be able to provide enough information to select the single best model. Therefore, integration of such models with equivalent performance may be a better choice. Secondly, ensembles can compensate for the imperfection in individual search process. Thirdly, in real practice, there may not be a true target function at all. Ensembles can provide a relatively good approximation and hence, result in a better generalization performance. Ensemble learning has been used in various applications, such as face recognition, medical diagnosis, and gene expression analysis [30].

In general, ensemble learning can be summarized as a two-step approach. The first step aims to develop a number of base models either in a parallel or in a sequence way. The base models developed in a parallel way have little connections between each other. By contrasts, if base models are developed in a sequence way, one base model has influence on the generation of its subsequent base model. The second step uses base models to output the final result through certain combination schemes, such as the majority voting and weighted average methods. To construct good ensemble models, there are two main considerations. Firstly, the performance of base models should be as more accurate as possible [30]. Secondly, the contained base learners should be as more diverse as possible. The diversity of base learners can be introduced in several ways, such as manipulating training samples, manipulating input variables, manipulating output variables, injecting randomness into learning algorithm, and adopting multiple learning algorithms [30].

2.7. Performance evaluation indices

Two sets of performance indices are used in this study. The first set uses root mean square error (RMSE) and R^2 for the input selection and model parameter optimization. RMSE is a scale-dependent metric and it results in values with the same units of the measurements. R^2 is the coefficient of determination, which ranges from 0 to 1, reflecting the goodness-of-fit. The definitions of these two metrics are shown in Eqs. (4) and (5).

$$\mathrm{RMSE} = \sqrt[2]{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n}} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} \tag{5}$$

The second set of indices adopts RMSE, mean absolute error (MAE) and mean absolute percentage error (MAPE) to evaluate prediction accuracies. MAE is a scale-dependent metric, which effectively reflect the prediction error by preventing the offset between positive and negative errors. MAPE is a scale-independent metric, offering a more straightforward way to describe the accuracy. The definitions of these two metrics are shown in Eqs. (6) and (7).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}\left|Y_i - \widehat{Y}_i\right|}{n} \tag{6}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|Y_i - \widehat{Y}_i\right|}{Y_i} \tag{7}$$

where Y_i is the actual measurement; $\widehat{Y}_i$ is the predicted value; n is the number of measurements.

3. Development of prediction models

3.1. Data preparation

The data set adopted in this study can be classified into two categories. The first category contains the building power consumption data, collected at 15-min intervals from the International Commerce Center (ICC) [32]. ICC is a high-class skyscraper with a height of approximately 490 m and a floor area of 321,000 m^2, which is the tallest building in Hong Kong currently. ICC consists of a 4-floor basement (24,000 m^2) mainly used for parking, a 6-floor block building (67,000 m^2) used for hotel ballrooms and shopping arcades, and a 98-floor tower (230,000 m^2) for commercial offices and a six-star hotel. ICC has been put into operation since 2010. ICC is a typical high-rising high-class commercial building in Hong Kong. One-year (i.e., 2011) data are used for this study, resulting in around 34,616 observations in total. The second category includes the meteorological data of the year 2011 obtained from the Hong Kong Observatory. The data were recorded daily. In total, twelve meteorological variables are considered and summarized in Table 1. Five variables, which describe the time of observations, i.e., month, day, weekday, hour, and minute, are also included. Two kinds of models are developed to predict the next-day daily energy consumption (i.e., Model-1) and peak power demand (i.e., Model-2). These models are suffixed by "-1" and "-2", respectively. For instance, the model developed for next-day daily energy consumption using SVR is denoted as "SVR-1".

3.1.1. Data transformation

Two types of data transformation are conducted in this study. The first one was performed prior to the feature extraction to remove the embedded seasonal effect in building energy consumption data. Such seasonal effect may negatively affect the performance of the following mining steps, e.g., outlier detection. For instance, in Hong Kong, the building's daily power consumption in summer are much larger than those in winter, since cooling is needed in summer while heating is usually not provided in winter. Consequently, it is very difficult to identify abnormal profiles using raw data directly. To overcome this problem, standardization is applied to each day. The power consumption values in each day

Table 1
Summary of meteorological variables.

Meteorological variables	Units	Range
Maximum dry-bulb temperature	°C	[10.4, 35.0]
Mean dry-bulb temperature	°C	[8.8, 30.9]
Minimum dry-bulb temperature	°C	[7.2, 28.8]
Mean dew point temperature	°C	[−0.7, 26.2]
Mean relative humidity	%	[39.0, 97.0]
Mean pressure	hPa	[997.8, 1025.6]
Mean amount of cloud	%	[3.0, 100.0]
Total rainfall	mm	[0.0, 106.0]
Number of hours of reduced visibility	h	[0.0, 12.2]
Daily global solar radiation	MJ/m^2	[1.1. 28.3]
Total evaporation	mm	[0.1, 7.9]
Mean wind speed	km/h	[5.0, 61.9]

are first centered by daily mean and then scaled by daily standard deviation. In this way, the seasonal effect is filtered out and the relative trend within each day is obtained.

The second type of data transformation aims to provide a suitable data set for input selection and model development. The reasons behind are twofold: firstly, the ranking criterion used for input selection is the square of variable weight. If the input variables are of different scales, the associated weights cannot be correctly reflect the true variable importance; secondly. The successful implementation of some predictive algorithms, e.g., SVR, requires the input variables to have similar scales. Otherwise, input variables with larger scales will dominate the modeling process. Therefore, all the candidate inputs, including the power consumption data and meteorological data, are standardized in this study.

3.1.2. Feature extraction

The raw data were collected at an interval of 15 min, resulting in 96 values for each daily energy consumption profile. Such high dimension representation of data is not efficient as these values are normally correlated. Usually, high dimensional data lead to a higher computation load and worse performance if distance-based techniques are used. In this study, a simple but efficient way of feature extraction is proposed to remove redundant information. Based on domain knowledge, it is realized that commercial building usage patterns are relatively fixed. It is feasible to represent the daily energy consumption profiles using summary statistics in different time periods.

More specifically, each daily profile is divided into three modes, i.e., morning mode (07:00–13:00), afternoon mode (13:00–20:00), and night mode (20:00–07:00). For each mode, 4 features are extracted by calculating the mean, maximum, minimum, and standard deviation of the standardized daily time series. As a result, 12 features are extracted to represent the daily energy consumption profile.

3.1.3. Creation of candidate input pool

Based on the result of feature extraction, the raw building energy consumption data are now transformed to multiple time series data representing 12 features. Combined with meteorological data and output variables, each daily observation has 29 variables, i.e., month, day, weekday, 12 extracted features, 12 meteorological variables, total energy consumption, and peak power demand. To create a pool of candidate inputs, an exploratory study is conducted using autocorrelation function (ACF) and partial autocorrelation function (PACF). These two tools are commonly used for time series data to identify repeating patterns or select inputs for modeling. ACF is the linear dependence of a variable with itself at two points in time. PACF is the autocorrelation between two points in time without considering the linear dependence of observations between these two time points. The ACF and PACF of daily energy consumption and daily peak power demand are investigated and all the variables with time lags considered to be significant are included in the pool of candidate inputs. The significance level is selected as 10%. The maximum time lag considered is 25 days.

The resulting ACFs and PACFs for daily energy consumption and daily peak power demand are shown in Figs. 2 and 3, respectively. The blue dotted lines show the significance threshold and any correlation value exceeding such threshold is considered significant. The ACFs for both outputs show an obvious repeating pattern of seven days. It is observed that the PACFs with time lags of more than 15 days are not significant any more. Hence, in this study, all the variables with a time lag of up to 15 days are included in the input candidate pool, from which the RFE is used to draw inputs. As a result, the candidate input pool has 450 variables in total, containing the variables of month, day, weekday, 12 meteorological variables of the prediction day, and all the variables with time lags up to 15 days (i.e., $29 \times 15 = 435$).

3.2. Identification of abnormal daily building energy consumption profiles

The abnormal building energy consumption profiles are profiles having significant difference with normal profiles. Such profiles should be identified and removed prior to model development as they may negatively affect the performance of predictive algorithms. The identification of abnormal daily profiles can be regarded as an outlier detection problem. To detect outliers, it is essential to construct a reference set for each observation. In this study, the EWKM clustering algorithm is first applied to the extracted features to identify typical daily profiles, and therefore, constructing reference sets for outlier detection. Then, the GESD algorithm is applied to the twelve extracted features. Daily profiles with features being suspected as outlier are regarded as abnormal profiles.

As introduced in Section 2.2, the implementation of EWKM algorithm requires the definition of three parameters, i.e., the cluster number (k), the weight distribution parameter (λ), and the convergence threshold (δ). As recommended in [19], λ should be constrained between 1 and 3, and δ is taken as constant as 0.00001. The parameters k and λ are determined by optimizing the clustering performance. Two internal cluster validity measures, i.e., the Dunn index and Davis–Bouldin index, are used to indicate the cluster performance. The Dunn index, which defines the ratio between the minimal intracluster distance to maximal interclsuter distance, ranges from 0 to infinity and it should be maximized. The Davies–Bouldin index, which is the ratio of the sum of intracluster scatter to interclsuter separation, ranges from 0 to infinity and it should be minimized. The candidate values for k and λ ranges from 2 to 10 and 1 to 3, respectively. The score of the objective function is defined as the difference between the Dunn index and Davies–Bouldin index, and it should be maximized. The optimization process is shown in Fig. 4. It shows that the optimized k and λ values are 2 and 2.65, respectively. The resulting clustering performance is shown in Fig. 5. It is observed that almost all the observations in cluster No. 1 come from weekdays and nearly all the observations in cluster No. 2 are from weekends. The clustering results meet the domain knowledge, as the operating patterns for commercial buildings are quite different between weekdays and weekends due to the great change in occupancy. The results indicate that the outlier detection should be performed on two clusters separately.

Two parameters of the GESD algorithm, i.e., N and α are defined as half of the observation number in clusters and 5%, respectively. The GESD algorithm is applied repeatedly for each feature in each cluster. Observations with features being identified as outliers are regarded as abnormal. The results are summarized in Table 2. In total, 18 out of 347 observations were identified as outliers. More specifically, 5.67% and 4.00% of the observations were identified as outliers for the weekday data (i.e., cluster No. 1) and weekend data (i.e., cluster No. 2), respectively. It is observed that the identified outliers can be categorized into two types, i.e., public holidays excluding Sundays, and the days adjacent to those public holidays. 14 out of 20 public holidays, which took place in 2011, appear as outliers. The other 4 of 18 outliers identified are those days adjacent to special public holidays. Again, the results meet common sense, as the occupancy tends to drop greatly for special public holidays or days in adjacent to them. Figs. 6 and 7 are shown as examples of two kinds of abnormal profiles. More specifically, Fig. 6 shows the abnormal power consumption profiles of April 5, 2011 (i.e., the Ching Ming Festival) and compares it with the profile

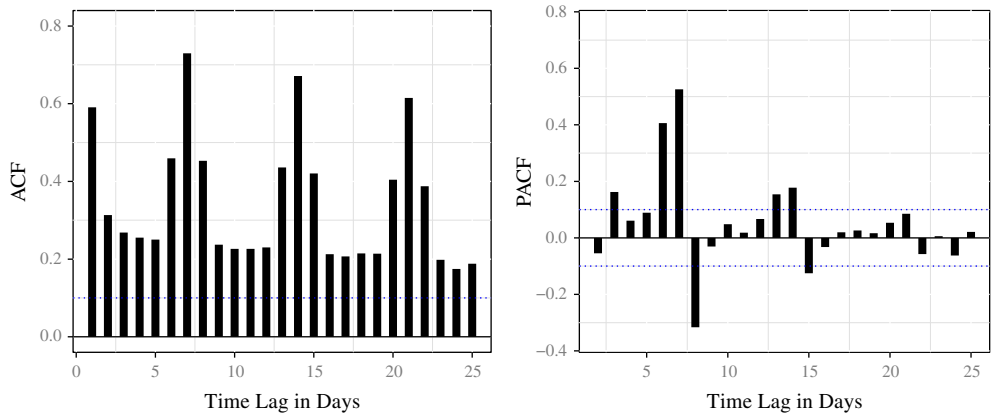

Fig. 2. ACF and PACF of daily energy consumption.

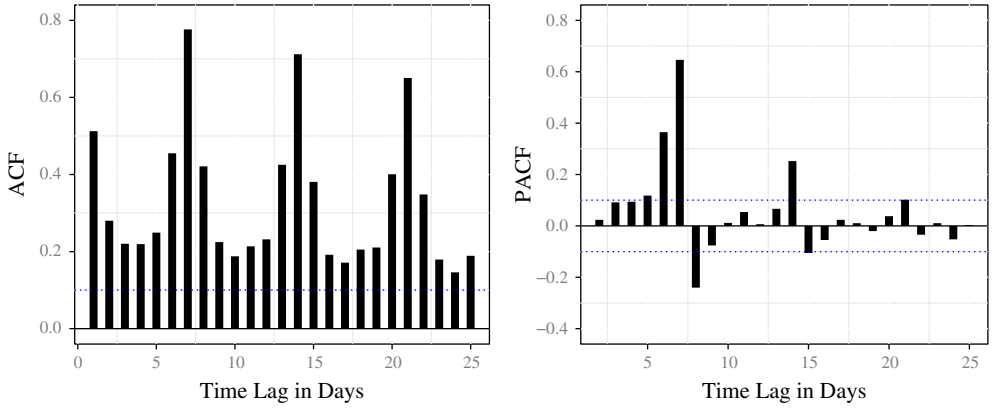

Fig. 3. ACF and PACF of daily peak power demand.

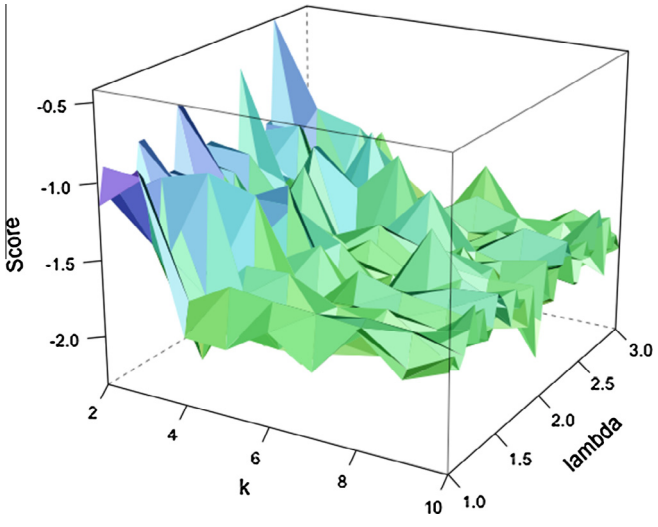

Fig. 4. Optimization process of EWKM parameters k and λ.

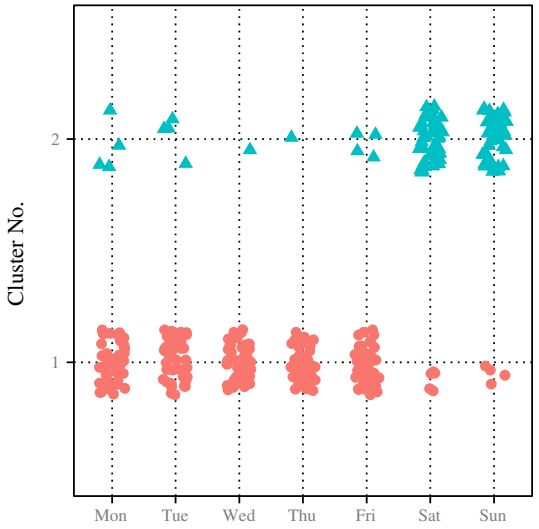

Fig. 5. EWKM clustering performance.

of 7 days later. The identified abnormal energy consumption profile consume much less energy than those of normal working days. Fig. 7 shows the abnormal consumption profile of October 6, 2011, which is the day after one special public holiday (i.e., October 5, 2011 is the Chung Yeung Festival) and compares it to that of a normal weekday 7 days later. Compared to the first example, the difference here is not that significant, but still, a clear deficit can

Table 2
Summary of outlier detection.

Description	Total observations	Number of outliers	Outlier percentage (%)
Weekdays	247	14	5.67
Weekends	100	4	4.00
Total	347	18	5.19

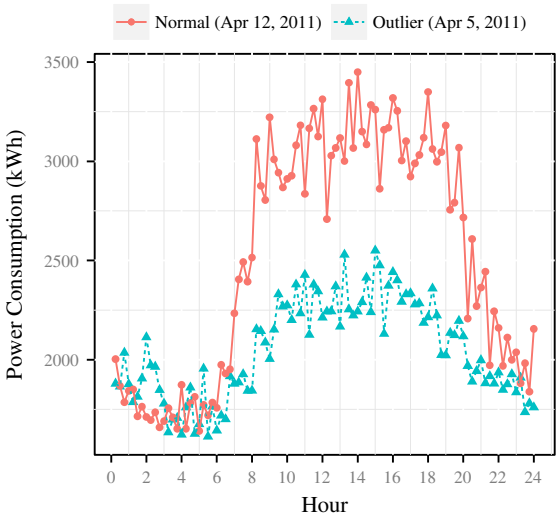

Fig. 6. Abnormal profiles of one public holiday.

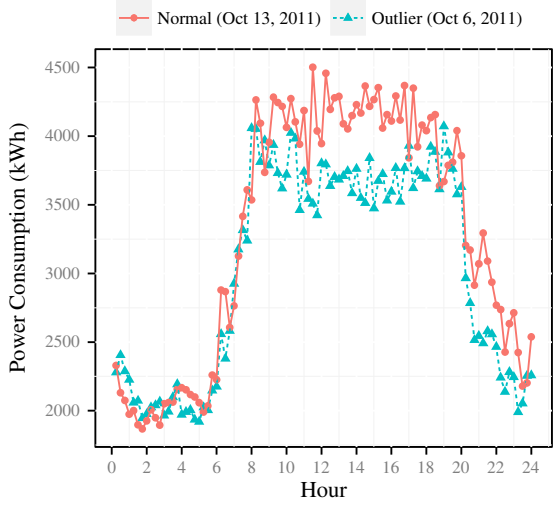

Fig. 7. Abnormal profiles of one day adjacent to public holiday.

be observed. It is realized that the inclusion of such abnormal observations may negatively affect the modeling process of predictive algorithms and hence, should be detected and removed.

3.3. Input Variable Selection

The RFE is used to perform the input variable selection process for seven predictive algorithms, i.e., MLR, SVR, RF, MLP, BT, MARS, and *k*NN. The other predictive algorithm, ARIMA, is a univariate modeling algorithm which only uses the output itself to construct the models and hence, no input variable selection process is needed. The whole data set was divided into two parts, i.e., training (70% of all data sets) and testing (30%) data sets. Three breaks were defined on the quantiles of the considered output and stratified random splits were performed within these sub-sections. The input variable selection was performed based on the training data set. The size of inputs under evaluation was defined as a vector, containing integers from 5 to 40. To obtain unbiased evaluation, bootstrapping was selected as the resampling method and the resampling was conducted 200 times. The model performance was evaluated using the rest 30% of data set. The subset of candidate inputs, which resulted in the best prediction performance, was selected for model development. The first set of performance evaluation metrics, i.e., RMSE and R^2, were used in this stage.

The optimal numbers of inputs for different predictive algorithms are summarized as follows. For the prediction models of next-day energy consumption, the numbers of inputs are 24, 29, 31, 36, 33, 12, and 27 for MLR, SVR, RF, MLP, BT, MARS, and *k*NN, respectively. The numbers are 27, 30, 34, 35, 33, 15, and 30 for the prediction models of next-day peak power demand. As an example, Fig. 8 illustrates the RFE process for RF-1. The black points represent the averaged performance indicators and the error bars present the standard errors during the resampling process. The best cases are highlighted using the red crosses. It is observed that the minimum RMSE and maximum R^2 can be achieved when the input numbers are 31. The variable importance for RF-1 and RF-2 is shown in Figs. 9 and 10. It is shown that the peak power demand and daily energy consumption of 7 days and 14 days prior to the prediction day are the top 4 most important inputs for both RF-1 and RF-2. It meets the domain knowledge, as the energy consumption patterns of commercial day on the same weekday are quite similar to each other. Some input variables are commonly selected for different algorithm, e.g., the

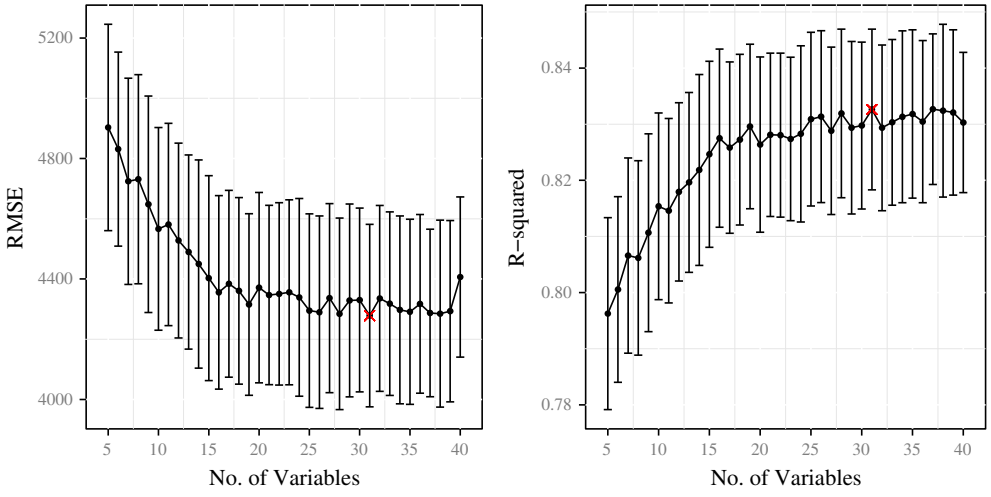

Fig. 8. Variable selection process for RF-1.

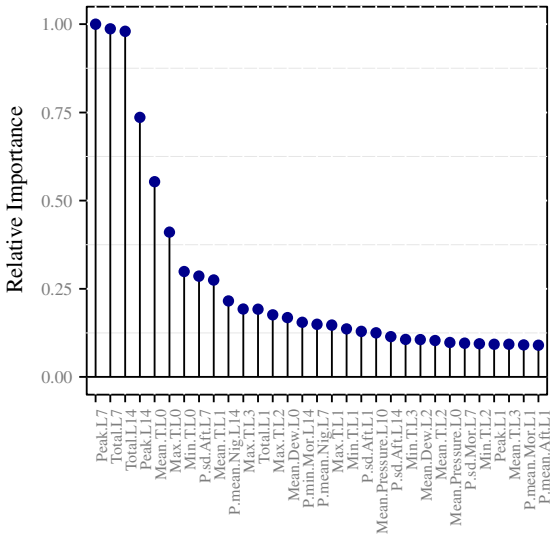

Fig. 9. Variable importance for RF-1.

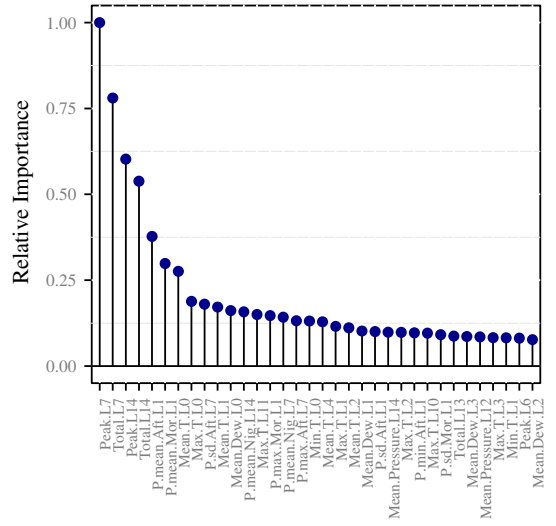

Fig. 10. Variable selection process for RF-2.

maximum dry-bulb temperature of the prediction day, and the extracted building energy consumption features of the day before prediction day. The selection of such variables is reasonable. For instance, the maximum dry-bulb temperatures of the prediction day should also have high ranks, as the building energy consumption is closely related to the outdoor weather condition. It is noted that some meteorological variables, which have smaller correlations with the outputs, e.g., the atmosphere pressure, are sometimes selected. By contrast, some meteorological variables, e.g., the relative humidity and wind speed, which were considered as important variables in other research work [14,15], are not selected for all 7 algorithms, even though they have higher correlations with the outputs. As indicated by Guyon and Elisseeff [25], the variables selected by RFE may not be the ones that are individually most relevant to the output, but as a whole, they form the most compact and useful subset for modeling. The employment of RFE enables an objective way of input selection. Little domain knowledge is needed. The RFE algorithm is able to automatically pick the optimal input combination for different predictive algorithms from different data sets, resulting in more flexibility in real applications.

3.4. Optimization of model parameters

Based on the RFE results, prediction models were developed using different algorithms. Optimization of model parameters was performed based on the LGOCV technique. The LGOCV repeatedly splits the data into training and testing data sets. The second set of evaluation metrics, i.e., RMSE and R^2, are used as performance indicators. In this study, the resampling number was set to be 50. During each resampling round, 70% of the entire data were used for model training and the performance was evaluated using the rest 30% of data.

The optimization results of other predictive algorithms are summarized as follows. No parameter needs to be optimized for MLR and the coefficients were fit through maximum likelihood estimation. Two parameters, i.e., the regularized constant (i.e., C) and the inverse width parameter of Gaussian radial basis function (i.e., sigma), are optimized for SVR. The optimal C values were 1.0 for SVR-1 and 0.5 for SVR-2. The optimal sigma values were 0.048 for SVR-1 and 0.050 for SVR-2, respectively. The model parameter to be optimized for RF is the number of random inputs to be considered when splitting each node, and the optimized values were 5 and 10 for RF-1 and RF-2, respectively. For MLP, the parameter to be optimized is the size of hidden neurons, and the results were 20 and 14 for MLP-1 and MLP-2, respectively. Two parameters, i.e., the maximum tree depth and the number of initial boosting iterations, should be optimized for BT. The optimal maximum tree depths were determined as 3 and 5 for BT-1 and BT-2, respectively. The optimal number of initial boosting iterations was 150 for both BT models. The maximum number of terms in pruned model and the maximum degree of interaction were optimized for MARS. The corresponding values for both MARS models were 10 and 1. The value k should be optimized for kNN models and the optimized value was 7 for both models. Three parameters, which define the numbers of autoregressive terms, differencing order, and moving average terms, were optimized for ARIMA models. The corresponding values are (15, 1, 6) and (10, 1, 2) for ARIMA-1 and ARIMA-2, respectively.

3.5. Development of ensemble models

Each ensemble model to be developed combines eight base models. The weights associated with each base model are determined using the genetic algorithm (GA). The objective function is defined to minimize the MAPE. The weight of individual base model was constrained between 0 and 1. The weights were normalized in such a manner that the total sum of weight equals to 1. To obtain mature optimization results from GA, the population size, number of iterations, mutation chance, and the proportion of elitism were set to be 250, 1000, 0.1, and 0.2, respectively. The results are summarized in Table 3.

4. Results and discussion

4.1. Prediction performance

Three evaluation metrics, i.e., RMSE, MAE, MAPE, are used to evaluate the performances of the prediction models. The performances of the eight base prediction models and the ensemble models are summarized in Tables 4 and 5.

For the two prediction cases, SVR and RF produce the most accurate results. These two techniques have been widely applied owing to their effectiveness in mapping complex nonlinear relationship. The results show that they are effective in predicting building energy consumption. ARIMA results in the worst performance, which is reasonable considering that the ARIMA model

Table 3
Weights of base models.

Base models	Model-1	Model-2
MLR	0.087	0.109
ARIMA	0.008	0.060
SVR	0.315	0.217
RF	0.404	0.381
MLP	0.076	0.076
BT	0.066	0.047
MARS	0.023	0.046
*k*NN	0.021	0.064

Table 4
Performances of the prediction models for the next-day energy consumption.

Models	RMSE	MAE	MAPE (%)
MLR-1	5585.47	4406.95	4.23
ARIMA-1	6090.31	5131.32	5.45
SVR-1	4196.01	3016.83	3.11
RF-1	4254.19	3163.64	3.17
MLP-1	5846.25	4840.10	4.75
BT-1	5167.15	4192.59	4.07
MARS-1	4951.90	4084.19	3.97
*k*NN-1	5046.06	4131.47	4.01
Ensemble-1	3393.27	2231.32	2.32

Table 5
Performances of the prediction models for the next-day peak power demand.

Models	RMSE	MAE	MAPE (%)
MLR-2	407.77	327.13	6.08
ARIMA-2	441.52	390.40	8.74
SVR-2	282.80	207.19	3.34
RF-2	295.69	216.35	3.63
MLP-2	431.04	338.39	6.46
BT-2	398.42	305.89	5.61
MARS-2	407.29	311.74	5.94
*k*NN-2	415.23	309.22	5.77
Ensemble-2	215.76	163.29	2.85

only uses the historical energy data as the inputs. Building energy consumptions are influenced by multiple factors, such as weather conditions and occupant numbers. Both ARIMA and MLR models are linear models. The use of multiple influential variables as inputs can improve the prediction accuracy as shown in the results of the MLR models. The MARS method, which extends the capability of MLR in modeling nonlinearity, does provide more accurate predictions. Therefore, nonlinearity does exist in the building energy consumption data, and the linear methods can hardly produce accurate results. The MLP method, which has been successfully used in many other fields, does not perform well in this study. The MLP models result in the second worst performance in both cases. It may because of the overfitting of the neural network models. The input numbers of the neural models are the largest. The other three predictive algorithms, i.e., BT, *k*NN and MARS, provide similar good performance. The research results show that ensemble models can achieve the best performance, with MAPEs of 2.32% and 2.83% for predicting the next-day energy consumption and peak load demand, respectively.

It is noted that the results presented in Tables 4 and 5 show a good consistency with the results presented in Table 3. In general, models with the higher accuracies tend to contribute more to the final ensemble model. For instance, the RF and SVR models are the most accurate and hence, they are assigned with the top two largest weights. By contrast, the ARIMA models receive the smallest weight, as their accuracies are the lowest. Nevertheless, it is not recommended to exclude those models with less encouraging results from the ensemble model. An ensemble model can achieve good performance even though the performance of its individual base models is poor [30]. The individual predictive algorithms have their own advantages. The merit of ensemble learning is that it can take most of the advantages of the base models to achieve the most accurate results. In addition, the errors of the base models are most likely uncorrelated and can be cancelled out in the ensemble model. Therefore, maintaining reasonable diversity among the base models is critical to the performance of the ensemble model.

Table 6
Required computation time for the proposed approach.

Performed task	Required computation time (s)			
	RFE-1	Model-1	RFE-2	Model-2
MLR	10.7	2.2	8.9	1.7
ARIMA	NA	8.0	NA	7.3
SVR	429.2	14.4	450.2	13.8
RF	101.7	28.6	108.8	35.3
MLP	2916.9	302.0	2809.3	332.4
BT	2037.3	19.2	2160.9	17.8
MARS	100.7	11.2	115.2	10.7
*k*NN	347.1	7.7	379.1	6.9

4.2. Computation efficiency

The computation of this study was performed on a Macintosh computer, OS version of 10.6.8, a processor of 2.2 GHz (Intel Core i7) and a memory size of 8 GB. The statistical programming language *R* [31] is used to perform all the computing tasks in this study.

The required computation time for two main steps, i.e., RFE, model training and optimization, is summarized in Table 6. Model 1 represents the prediction models for next day energy consumption, and Model 2 represents of the prediction models for next day peak demand. Once a model has been developed, the time spent for making prediction is so short that it can be neglected. Another main step is to optimize the weights of base models using GA. The computation took around 10 s for both ensemble models. In general, the RFE is the more computationally expensive step. The time taken by RFE is usually affected by the algorithm complexity. For instance, the development of MLR model was relatively easy and the corresponding RFE took the least time. By contrast, the development of the MLP models were complicated and the corresponding RFE processes were time-consuming.

5. Conclusions

Building energy consumption prediction is often needed in the evaluation of building performance, optimization of building operation, fault detection and diagnosis and demand side management for smart grid. This paper proposed a data mining-based approach to developing the ensemble models for predicting the next-day energy consumption and peak power demand. The ensemble models combine eight base models which are developed separately using eight popular prediction algorithms. Since each prediction algorithms have their pros and cons, the ensemble models enable the base models to complement with each other and hence produce better generalization performance. The approach is adopted to analyze the large energy consumption data of the tallest building in Hong Kong. The results show that the accuracies of the ensemble models are evidently better than those of base models. Considering the eight basis models, the support vector regression (SVR) and random forests (RF) models result in the best performance and therefore have the largest weights in the ensemble models. By contrast, traditional statistical methods, i.e., the multi-

ple linear regression (MLR) and ARIMA models, do not perform very well, because the building related processes are usually nonlinear and complex. Therefore, they have smaller weights in the ensemble models.

This paper also addresses the two major issues related to the development and application of the ensemble models, i.e., the data quality and the computation load. Through the use of feature extraction, clustering analysis, and GESD, abnormal daily energy consumption profiles are detected successfully. Most of the outliers identified come from public holidays excluding sundays and the days adjacent to those public holidays. The major reason is that the occupant number is very different from those on normal days. Such abnormal profiles are removed from the data set for the benefit of model development. It is shown that RFE process can properly select the optimal input variables, leading to reduced computation load and improved prediction performance. The computation times of the major steps are also presented in this paper. The model development processes are time consuming; however, the time for making prediction based on new inputs is very short once the models are ready. Therefore, the approach is applicable in practice.

Acknowledgement

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

References

[1] International Energy Agency (IEA), <http://www.iea.org/aboutus/faqs/energyefficiency/>, [accessed on October 2, 2013].
[2] Hong Kong energy end-use data 2012, Hong Kong Electrical & Mechanical Services Department; 2012.
[3] Li XL, Bowers CP, Schnier T. Classification of energy consumption in buildings with outlier detection. IEEE Trans Ind Electron 2010;57(11):3639–44.
[4] Potter CW, Archambault A, Westrick K. Building a smarter smart grid through better renewable energy information. In: Proceedings of IEEE/PES Power Systems Conference and Exposition, 15–18 March, Seattle, Washington D.C.; 2009.
[5] Hahn H, Meyer-Nieberg S, Pickl S. Electric load forecasting methods: tools for decision making. Eur J Oper Res 2009;199(3):902–7.
[6] Gonzalez-Romera E, Jaramillo-Moran MA, Carmona-Fernandez D. Monthly electric energy demand forecasting based on trend extraction. IEEE Trans Power Syst 2006;21(4):1946–53.
[7] Abdel-Aal RE. Modeling and forecasting electric daily peak loads using abductive networks. Int J Electr Power Energy Syst 2006;28(2):133–41.
[8] Yao R, Steemers KA. A method of formulating energy load profile for domestic buildings in the UK. Energy Build 2005;36(6):663–71.
[9] White JA, Reichmuth R. Simplified method for predicting building energy consumption using average monthly temperatures. In: Proceedings of the 31st intersociety energy conversion engineering conference, vol. 3; 1996. p. 1834–9.
[10] Aydinalp-Koksal M, Ugursal VI. Comparison of neural network, conditional demand analysis and engineering approaches for modeling end-use energy consumption in residential sector. Appl Energy 2008;85(4):271–96.
[11] Ma Y, Yu J, Yang C, Wang L. Study on power energy consumption model for large-scale public building. In: Proceedings of the 2nd international workshop on intelligent systems and applications; 2010. p. 1–4.
[12] Zhao HX, Magoules F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16(6):3586–92.
[13] Zhou Q, Wang S, Xu X, Xiao F. A grey-model of next-day building thermal load prediction for energy-efficient control. Int J Energy Res 2008;32(13):1418–31.
[14] Moazzami M, Khodabakhshian A, Hooshmand R. A new hybrid day-ahead peak load forecasting method for Iran's National Grid. Appl Energy 2012;101(1):489–501.
[15] Yalcintas M, Akkurt S. Artificial neural networks applications in building energy predictions and a case study for tropical climates. Int J Energy Res 2005;29(10):891–901.
[16] Dong B, Cao C, Lee SE. Applying support vector machines to predict building energy consumption in tropical region. Energy Build 2005;37(5):545–53.
[17] Solomon DM, Winter RL, Boulanger AG, Anderson RN, Wu LL. Forecasting energy demand in large commercial buildings using support vector machine regression. Department of Computer Science, Columbia University, Technical Report, CUCS-040-11; 2011.
[18] Magoules F, Zhao HX, Elizondo D. Development of an RDP neural network for building energy consumption fault detection diagnosis. Energy Build 2013;62(18):133–8.
[19] Jing LP, Ng MK, Huang JZ. An entropy-weighting *k*-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans Knowl Data Eng 2007;19:1026–41.
[20] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer Series in Statistics; 2009.
[21] Barnett V, Lewis T. Outliers in statistical data. 3th ed. New York: John Wiley & Sons Inc.; 1994.
[22] Rosner B. Percentage points for a generalized ESD many-outlier procedure. Technometrics 1983;25(2):165–72.
[23] Iglewicz B, Hoaglin DC. How to detect and handle outliers. 1st ed. Milwaukee: ASQ Quality Press; 1993.
[24] Seem JE. Using intelligent data analysis to detect abnormal energy consumption in buildings. Energy Build 2007;39(1):52–8.
[25] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.
[26] Shumway RH, Stoffer DS. Time series analysis and its applications with *R* examples. 3rd ed. New York: Springer; 2011.
[27] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[28] Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol 2008;77:802–13.
[29] Friedman JH. Multivariate adaptive regression splines. Ann Stat 1991;19(1):1–67.
[30] Dietterich TG. Ensemble methods in machine learning. In: Proceedings of the first international workshop on multiple classifier systems, Springer; 2000. p. 1–15.
[31] R Core Team. R: A language and environment for statistical computing. R Foundation Computing, Vienna, Austria. URL: <http://www.R-project.org/>.
[32] Ma ZJ. Online supervisory and optimal control of complex building central chilling systems. PhD thesis, The Hong Kong Polytechnic University; 2008.