Energy and Buildings 42 (2010) 1637–1646

Contents lists available at ScienceDirect

Energy and Buildings

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

A decision tree method for building energy demand modeling

Zhun Yu[a], Fariborz Haghighat[a,*], Benjamin C.M. Fung[b], Hiroshi Yoshino[c]

[a] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Quebec, Canada H3G 1M8
[b] Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, Canada H3G 1M8
[c] Department of Architecture and Building Science, Tohoku University, Japan

ARTICLE INFO

Article history:
Received 18 March 2010
Accepted 19 April 2010

Keywords:
Building energy consumption
Modeling
Decision tree
Classification analysis

ABSTRACT

This paper reports the development of a building energy demand predictive model based on the decision tree method. This method is able to classify and predict categorical variables: its competitive advantage over other widely used modeling techniques, such as regression method and ANN method, lies in the ability to generate accurate predictive models with interpretable flowchart-like tree structures that enable users to quickly extract useful information. To demonstrate its applicability, the method is applied to estimate residential building energy performance indexes by modeling building energy use intensity (EUI) levels. The results demonstrate that the use of decision tree method can classify and predict building energy demand levels accurately (93% for training data and 92% for test data), identify and rank significant factors of building EUI automatically. The method can provide the combination of significant factors as well as the threshold values that will lead to high building energy performance. Moreover, the average EUI value of data records in each classified data subsets can be used for reference when performing prediction. One crucial benefit is improving building energy performance and reducing energy consumption. Another advantage of this methodology is that it can be utilized by users without requiring much computation knowledge.

1. Introduction

There has been a growing concern about the total building energy consumption which is a substantial user of energy worldwide. Further, with rising living standards, building energy consumption throughout the world has been significantly increased over the past few decades. For example, from 1994 to 2004, building energy consumption in Europe and North America has increased at a rate of 1.5% and 1.9% per annum, respectively [1]. Chinese building energy consumption has increased at more than 10% per annum for the past 20 years [2]. The high level of building energy consumption and the steady increase in building energy demand necessitate designing energy efficient buildings and improving its energy performance.

In the practice of energy efficient building design, architects and building designers often need to identify which parameters will influence future building energy demand significantly. Furthermore, based on different combinations of these parameters as well as their values, architects and building designers usually expect to find a simple and reliable method to estimate building energy performance rapidly so that they can optimize their building design plans. Building energy simulation tools have been utilized to forecast and analyze building energy consumption and describe building energy use patterns, in order to benefit the design and operation of energy efficient buildings. In recent years, there have been many studies on building energy demand modeling and several methods were employed, such as traditional regression methods [3,4], artificial neural networks (ANN) methods [5–7], and building simulation methods [8,9], etc. Through statistical methods and regression equations, regression models correlate building energy demand with relevant climatic variables and/or building physical variables in order to predict energy demand. The main advantage of regression models is that they are comparatively simple and efficient. The ANN model is also able to predict the thermal performance of building and its foundation is based on mimicking the structure and properties of biological neural networks. The greatest strength of ANN models in comparison with other models lies in their ability to model complex relationships between inputs and outputs. These two methods have been successfully applied to predict building energy demand. However, considering the regression models are normally complicated equations and ANN models operate like a "black box"; therefore, the models developed using these methods are not understandable and interpretable especially for common users without advanced mathematical knowledge. This makes it difficult to be a common predictive tool. Moreover, in these studies, the focuses have been mainly on the energy use

* Corresponding author at: Department of Building, Civil and Environmental Engineering, Concordia University, 1455 De Maisonneuve Blvd., Montreal, Quebec, Canada H3G 1M8.
E-mail address: haghi@bcee.concordia.ca (F. Haghighat).

0378-7788/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.enbuild.2010.04.006

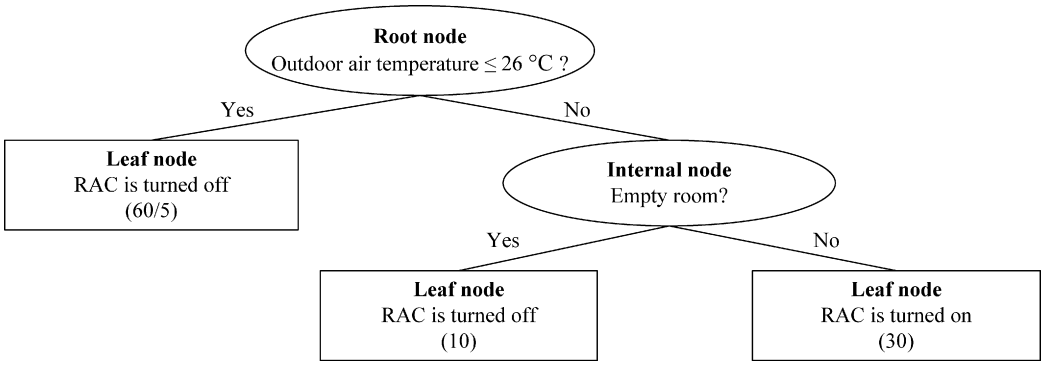

Fig. 1. Schematic illustration of a simple hypothetical decision tree.

prediction of existing buildings (e.g. predict hourly heating/cooling load for a certain type of building), whereas the energy use prediction of newly designed buildings, which is also very important for architects and building designers to make rational decisions at the early stage of design and operation, are seldom carried out.

Building simulation allows the prediction of building energy performance under various conditions. However, this method does not perform well in predicting the energy use for occupied buildings as compare to non-occupied buildings due to the lack of sufficient knowledge about occupants' behavior. Additionally, the application of building simulation programs is normally complicated and the learning process of these programs tends to be time-consuming.

In the past two decades, decision tree method, a novel computational modeling technique that uses flowchart-like tree structure, has been widely used for classification and prediction in many scientific and medical fields [10–12]. The popularity of decision tree method mainly attributes to its ease of use, and abilities to generate accurate predictive models with understandable and interpretable structures, which, accordingly, provide clear and useful information on corresponding domains. Moreover, the decision tree method is able to process both numerical and categorical variables, and perform classification and prediction tasks rapidly without requiring much computation efforts. However, it should be mentioned that decision tree method is more appropriate for predicting categorical variables than for predicting numerical variables. The application of decision tree method in building related studies is still very sparse. Tso and Yau [13] compared the accuracy of regression method, ANN method, and decision tree method in predicting average weekly electricity consumption for both summer and winter in Hong Kong. It was found that decision tree model and ANN model have a slightly higher accuracy than other models. Therefore, it is highly desirable to utilize decision tree method to process measured data, which has already included the influences of occupant activities, for building energy demand modeling.

The paper reports the development of a procedure to accurately estimate building energy performance indexes. The procedure is based on the decision tree method. The applicability of the procedure is then demonstrated for residential buildings sectors.

2. Methodology

2.1. Overview of decision tree

The decision tree methodology is one of the most commonly used data mining methods [14,15]. It uses a flowchart-like tree structure to segregate a set of data into various predefined classes, thereby providing the description, categorization, and generalization of given datasets. As a logical model, decision tree shows how the value of a *target variable* can be predicted by using the values of a set of *predictor variables*. Fig. 1 gives a decision tree indicating whether residents turn room air conditioners (RAC) on or off in their rooms in the cooling season. Assume 100 data records are used to build this decision tree and each record has three attributes: outdoor air temperature, room occupancy, and the operating state of RAC.

The target variable for the above decision tree is RAC operating states, with potential states being classified as either turning on or off. The predictor variables are outdoor air temperature ($\leq$26 °C or >26 °C) and room occupancy (empty or not). As shown in Fig. 1, the decision tree consists of three kinds of nodes: root node, internal node, and leaf node. Root node and internal node denote a binary split test on an attribute while leaf node represents an outcome of the classification and thus holding a categorical target label. Moreover, the numbers in the parentheses at the end of each leaf node depict the number of data records in this leaf. If some leaves are impure (i.e. some records are misclassified into this node), the number of misclassified records will be given after a slash. For example, (60/5) in the left most leaf in Fig. 1 means that, among the 60 records having outdoor temperature is lower than or equal to 26 °C that have been classified to *turned off*, 5 of them actually have the value *turned on*. By using this decision tree, whether RAC operating states should be classified as being 'turned on' or 'turned off' can be predicted. For example, if the outdoor air temperature is higher than 26 °C and the room is not empty, occupants will turn RAC on; otherwise they will turn it off.

2.2. Decision tree generation

Decision tree generation is in general a two-step process, namely learning and classification, as shown in Fig. 2. In the learning process, the collected data are split into two subsets, training set and testing set. Creation of training set and testing set is an important part of evaluating data mining models. Usually, most of the data records in the database are arbitrarily selected for training and the remained data records are used for testing. Note that training set and testing set should come from the same population but should be disjoint. Then, a decision tree generation algorithm takes the training data as input and outputs a decision tree. Commonly used decision tree generation algorithms include ID3 [14], classification and regression trees (CART) [16], and C4.5 [17]. In this study, we employ C4.5, along with an open-source data mining software WEKA, to build decision tree due to its flexibility and wide applicability to different types of data. In the classification process, the accuracy of obtained decision tree is first evaluated by making predictions against the test data. The accuracy of a decision tree is measured by comparing the predicted target values and the

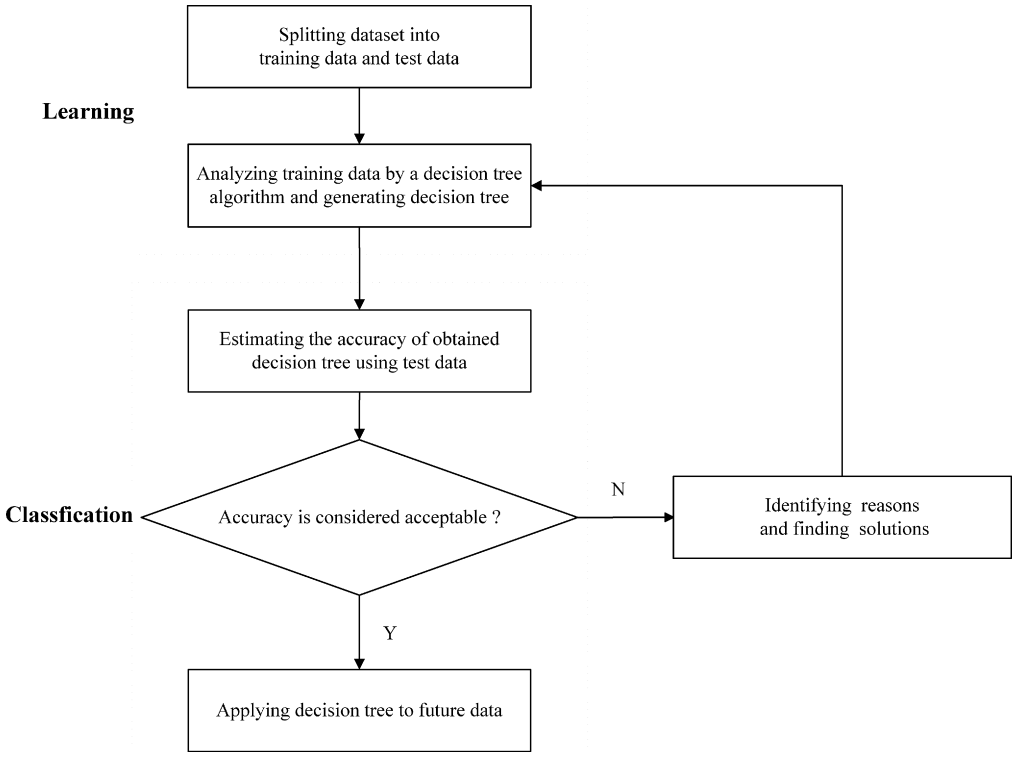

Fig. 2. Procedure of decision tree generation.

true target values of the testing data. If the accuracy is considered acceptable, the decision tree can be applied to new dataset for classification and prediction; otherwise, the reason should be identified and corresponding solutions should be adopted to tackle problems.

The procedure of generating a decision tree from the training data is explained as follows. Initially, all records in the training data are grouped together into a single partition. At each iteration, the algorithm chooses a predictor attribute that can "best" separate the target class values in the partition. The ability that a predictor attribute can separate the target class values is measured based on an attribute selection criterion, which will be discussed in Section 3.3. After a predictor attribute is chosen, the algorithm splits the partition into child partitions such that each child partition contains the same value of the chosen selected attribute. The decision tree algorithm iteratively splits a partition and stops when any one of the following terminating conditions is met:

1. All records in a partition share the same target class value. Thus, the class label of the leaf node is the target class value;
2. There are no remaining predictor attributes that can be used to further split a partition. In this case, the majority target class values become the label of the leaf node; and
3. There are no more records for a particular value of a predictor variable. In this case, a lead node is created with the majority class value in the parent partition.

2.3. Attribute selection criterion

The decision tree generation algorithm is a greedy algorithm. It iteratively splits a partition by choosing a split attribute that can best separate the target class values. The choice of split attribute determines the quality of the decision tree model and, therefore, the classification accuracy on the future data. The concept of *entropy* [16] in information theory is a widely criterion measure for decision tree to characterize the purity of a partition in decision tree nodes. Given a decision tree containing only binary target variables such as HIGH EUI and LOW EUI, the entropy of the data subset, D_i, of the ith tree node is defined as

$$\text{Entropy}(D_i) = -\left(\frac{{}^{n}\text{HIGH}}{\text{T\_N}}\log_2\frac{{}^{n}\text{HIGH}}{\text{T\_N}} + \frac{{}^{n}\text{LOW}}{\text{T\_N}}\log_2\frac{{}^{n}\text{LOW}}{\text{T\_N}}\right) \quad (1)$$

where ^{n}HIGH: the number of HIGH EUI records in D_i; ^{n}LOW: the number of LOW EUI records in D_i; T_N: the total number of records in D_i and T_N = ^{n}HIGH + ^{n}LOW.

The entropy varies between 0 and 1. Notice that the entropy equals to 0 if D_i is pure and it is 1 when ^{n}HIGH equals to ^{n}LOW. At each node of a decision tree, candidate splitting test will be used to evaluate all available attributes to select the most suitable attribute to split data. Suppose the jth attribute has been selected as node attribute. A candidate split test, ST, at the ith tree node is defined as

$$\text{ST} : \text{Val}_j(r) \le \text{T\_h} \quad (\text{if the } j\text{th attribute is numerical}) \quad (2)$$

$$\text{ST} : \text{Val}_j(r) \in \{v_1, v_2\}$$

$$(\text{if the } j\text{th attribute is categorical and has two values}) \quad (3)$$

where $\text{Val}_j(r)$: the value of the jth attribute of record r; T_h: threshold value; v_1, v_2: two values of the jth attribute.

Next, the algorithm applies ST to D_i and partitions D_i into two subsets, DS_1 and DS_2. Let r be a record in D_i. If the jth attribute is a numerical attribute, then

$$\text{DS}_1 = \{r \in D_i | \text{val}_j(r) \le \text{T\_h}\} \quad \text{and } \text{DS}_2 = \{r \in D_i | \text{val}_j(r) > \text{T\_h}\}. \quad (4)$$

If the jth attribute is a categorical attribute, then

$$\text{DS}_1 = \{r \in D_i | \text{val}_j(r) = v_1\} \quad \text{and} \quad \text{DS}_2 = \{r \in D_i | \text{val}_j(r) = v_2\}. \quad (5)$$

Let m and n be the number of records in DS_1 and DS_2, respectively. The entropy after the split test can then be calculated as

the weighted sum of the entropies for the individual subsets

$$\text{Entropy}(\text{DS}_1 \text{ and } \text{DS}_2) = \frac{m}{m+n}\text{Entropy}(\text{DS}_1) + \frac{n}{m+n}\text{Entropy}(\text{DS}_2) \quad (6)$$

where Entropy (DS_1) and Entropy (DS_2) can be calculated by using formula (1).

The selection of node attribute used to split data is very important and a rational selection can improve the purity of tree nodes. A widely used attribute selection measure is *information gain* [18], which is defined as the entropy reduction before and after a candidate splitting test. Therefore, information gain can be calculated as

$$\text{InfoGain} = \text{Entropy}(D_i) - \text{Entropy}(\text{DS}_1 \text{ and } \text{DS}_2) \quad (7)$$

For each tree node, the attribute with the maximum information gain will be chosen as the splitting attribute at this node. The information gain measure, however, has a bias to attributes with larger number of domain values. One way to avoid such bias is to normalize the information gain by a split information value defined analogously with information gain. C4.5 employs this improved measure, *gain ratio* [15]:

$$\text{GainRatio} = \frac{\text{InfoGain}}{\text{SplitInfo}} \quad (8)$$

where

$$\text{SplitInfo} = -\left(\frac{m}{m+n}\log_2\frac{m}{m+n} + \frac{n}{m+n}\log_2\frac{n}{m+n}\right) \quad (9)$$

The attribute with the highest gain ratio is selected as the splitting attribute.

Additionally, in order to detect whether a node should be a leaf, a minimum threshold value of entropy (EN_{min}) will be predefined and compared with node classification entropy (Entropy (D_i)), if Entropy (D_i) is lower than EN_{min}, then this node is a leaf and will be labeled LEAF. Otherwise a further splitting test should be performed. However, if no significant effects can be observed on information gain or gain ratio in further candidate splitting tests, the test will be also stopped and the node will be labeled STOP.

3. Data source and basic analysis

3.1. Data collection and pre-processing

To evaluate and improve residential building energy performance in Japan, a project was performed by Research Committee on Investigation on Energy Consumption of Residential Buildings (2001–2003) and Committee on Energy Consumption of Residential and Countermeasures for Global Warming (2004–2005) of the Architectural Institute of Japan. This analysis used the data base of Cd-Rom titled "Energy Consumption for residential buildings in Japan" [19]. In this project, field surveys on energy related data and other relevant information were carried out in 80 residential buildings located in six different districts in Japan.

- Energy end use of all kinds of fuel used by the building at different time intervals;
- Indoor environment parameters every 15 min;
- Household characteristics; and
- Other issues such as occupant behaviors and energy saving measures;

Fig. 3 shows the boxplot for monthly average outdoor air temperature in each district in 2003 using Japanese meteorological data. The mean value of monthly average temperature, i.e. annual average temperature, is also given. Clearly the monthly average temperature has a more or less symmetric distribution. The annual average temperature is higher than 8 °C in all the six districts and the temperature in Hokkaido and Tohoku is comparatively lower than other districts.

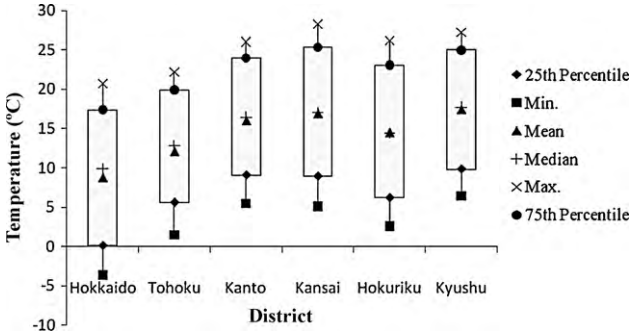

Fig. 3. Boxplot for monthly average outdoor air temperature in the six regions in 2003.

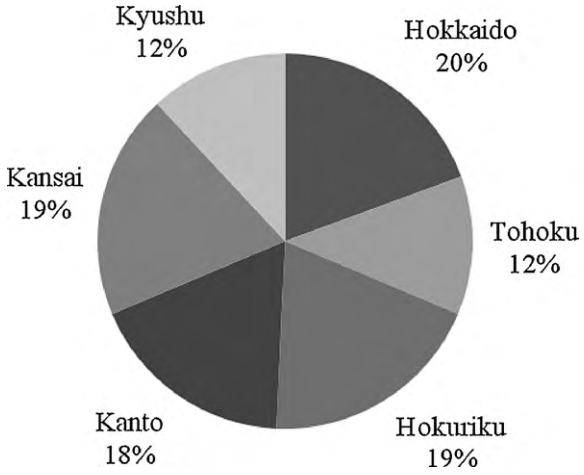

Fig. 4. Percentage breakdown.

Scrutinizing the data from the 80 buildings it was found that only 67 sets were complete while the other 13 had missing values of energy consumption data. Fig. 4 shows the percentage breakdown of available residential buildings in each district. It can be seen that the distribution is roughly uniform.

Data reduction and aggregation was also performed as a preprocessing step of preparing the data for a database. For example, the primary energy sources in the investigated families include electricity, natural gas, and kerosene. All these energy sources are converted into an equivalent energy value based on conversion coefficients in Table 1.

Moreover, energy end use is classified into eight categories and the three major categories include the space heating/cooling, hot water supply, and kitchen. Each end use data with interval of 5 min was aggregated so as to compute hourly, daily, monthly, and annual total amounts. And thus total energy use can also be calculated as the sum of the energy content of all the fuel used

Table 1
Conversion coefficients of different fuels.

Fuel	Conversion coefficient	Unit
Electricity	3.6	MJ/kWh
City gas (4A–7C)	20.4	MJ/Nm^3
City gas (12A–13C)	45.9	MJ/Nm^3
Liquefied petroleum gas (LPG)	50.2	MJ/Nm^3
Kerosene	36.7	MJ/L

Table 2
Summary of model inputs.

Number	Variable	Type	Value	Variable label (unit)
1	TEMP	Categorical	High/low	Annual average air temperature
2	HOUS	Categorical	Detached/apartment	House type
3	CONS	Categorical	Wood/non-wood	Construction type
4	AREA	Numerical	[70, 240]	Floor area (m^2)
5	HLC[a*]	Numerical	[1.01, 4.35]	Heat loss coefficient (W/m^2K)
6	ELA[b*]	Numerical	[0.35, 13.30]	Equivalent leakage area (cm^2/m^2)
7	NUM	Numerical	[2, 6]	Number of occupants
8	HEAT	Categorical	Electric/non-electric	Space heating
9	HWS	Categorical	Electric/non-electric	Hot water supply
10	KITC	Categorical	Electric/gas	Kitchen

[a*] Calculated based on building design plans.
[b*] Measured by the fan pressurization method.

by the building in a year. Based on above work, a database was created.

3.2. Model target variable

In order to demonstrate building energy performance, model target variable is expressed in energy use intensity (EUI), defined as the ratio of annual total energy use to total floor area (the annual total energy use is calculated as the sum of the energy content of all fuel used by the building in 2003). As mentioned previously, decision tree method is more appropriate for predicting categorical variables. Therefore, a concept hierarchy for building EUI is formed before classification and prediction are carried out. Due to the small database size, a two-grade descending scale, i.e. high level and low level, corresponding to low energy performance and high energy performance, are considered applicable and understandable. Building EUI ranges from 176 MJ/m^2 to 707 MJ/m^2 in the database and thus data ranged from the average of the maximum and minimum to the maximum value, i.e. [441.5, 707], is considered 'HIGH'. And data from the minimum value to the average of the maximum and minimum, i.e. [176, 441.5] is considered 'LOW'.

It should be mentioned that, decision tree can also be used to classify and predict multiple EUI levels rather than just two. For example, instead of 'HIGH' and 'LOW', a concept hierarchy of EUI may map real EUI values into four conceptual levels such as *EXCELLENT*, *GOOD*, *FAIR*, and *COMMON*, thereby resulting in a smaller data range of each level and providing a more detailed description. However, more conceptual levels require a larger database and may be prone to higher misclassification rate of data records and thus reduce the accuracy of decision tree models.

3.3. Model input variables

Ten parameters (or *attributes*) are selected from the database to be model input variables and the summary of these parameters is given in Table 2.

These ten parameters are grouped into four categories that are important determinants of household energy demand.

(1) Climatic conditions (TEMP). The range of annual average outdoor air temperature in the six districts is discretized into two intervals based on the same concept hierarchy as the EUI mentioned earlier: the high interval (8.8 °C, 13.1 °C), and the low interval (14.3 °C, 17.4 °C). According to this discretization criterion, the low temperature districts include Hokkaido and Tohoku while the other four districts belong to high temperature districts;
(2) Building characteristics (HOUS, CONS, AREA, HLC, ELA). For building construction type, the non-wood type includes steel reinforced concrete (SRC), reinforced concrete (RC), and steel structure (S);
(3) Household characteristics (NUM); and
(4) Household appliance energy sources (HEAT, HWS, KITC). Energy sources are divided into energy generated from electricity consumption and energy generated from other fuels such as kerosene and natural gas.

Fig. 5 shows the distribution of all the categorical parameters. It can be observed that all the percentages range from 30% to 70%, indicating a fairly uniform distribution.

4. Results and discussion

C4.5 algorithm was used for training data set (55 records were arbitrarily selected from the database) and test data set (i.e. the remained12 records that are independent of training set) by using WEKA to build a decision tree for predicting whether the EUI of residential buildings should be classified as being 'HIGH' or 'LOW'.

4.1. Generation of decision tree

Fig. 6 shows the decision tree for the classification of building EUI levels. This decision tree is built on the basis of the training data set of 55 data records with the ten attributes list of Table 2. It can be seen that this tree includes a total of 21nodes among which 11 are leaf nodes, including 8 LEAFs and 3 STOPs: this represents 11 classes (either EUI = HIGH or EUI = LOW). The explanatory note of three kinds of nodes, namely root node, internal node, and leaf node in this decision tree is shown in Fig. 7. Note that entropy is also calculated and given in each node to characterize the purity of the sub dataset in that node. Moreover, the average EUI value of data records in each class is given and used for reference when per-

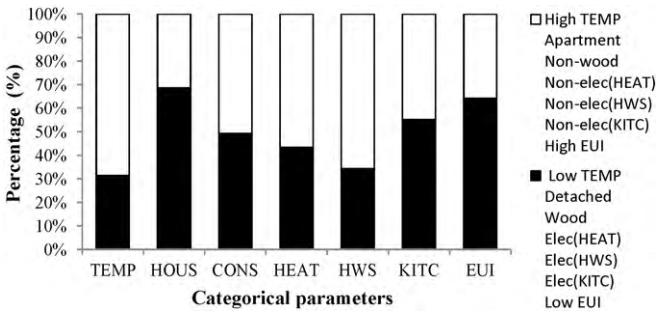

Fig. 5. Categorical distribution of the six categorical parameters.

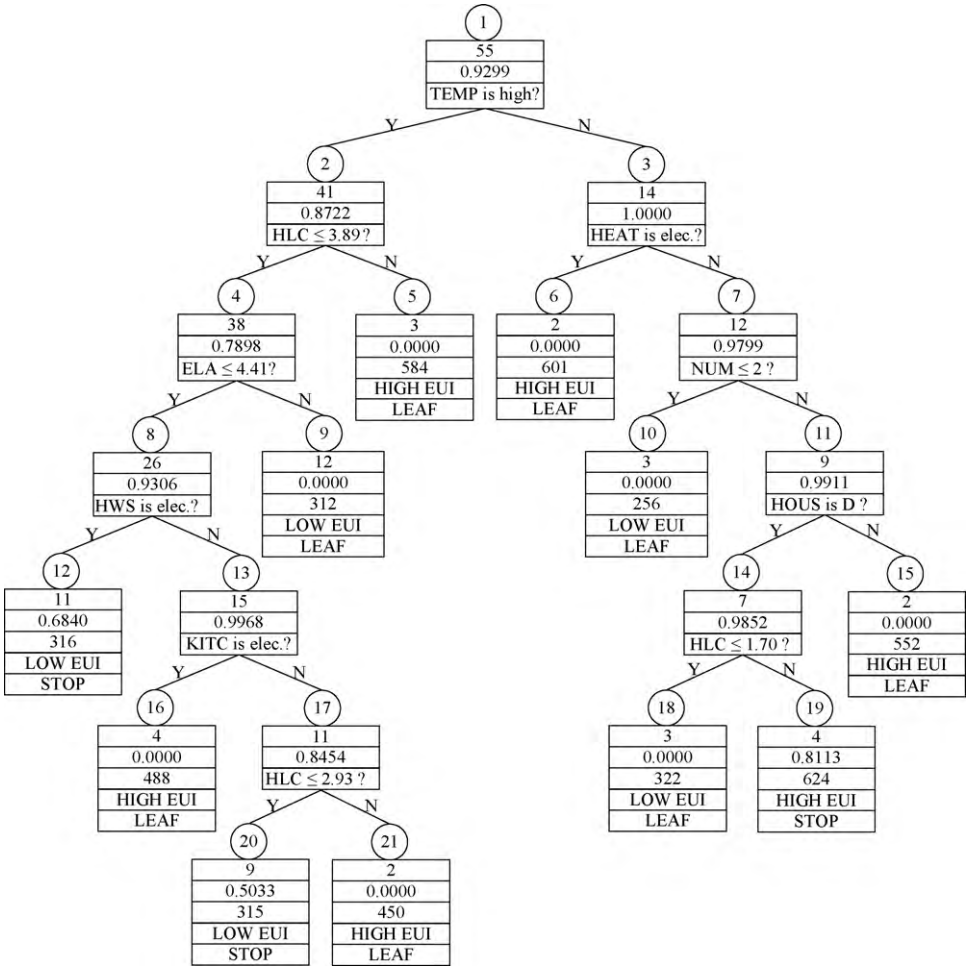

Fig. 6. Decision tree for the prediction of building EUI level.

forming prediction. Specifically, this reference value can be viewed as predictive numerical EUI value of the new data records that fall into that class.

The WEKA analysis report also provides the information on the classification accuracy of the decision tree. The report indicates that 51 records which accounts for 93% of all the training records are correctly classified: this indicates a good accuracy. Also, confusion matrix reports how many data records are correctly classified and misclassified in the class of HIGH EUI and LOW EUI separately, as below:

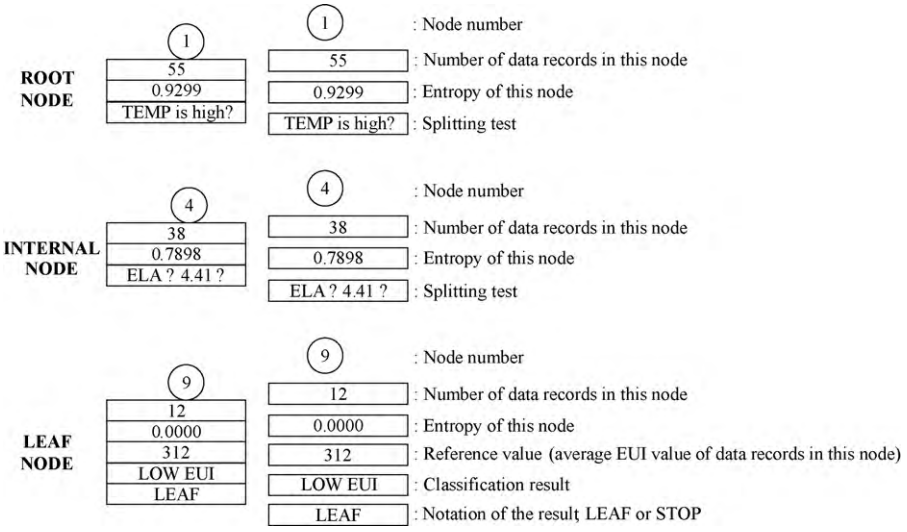

Fig. 7. Explanatory note of decision tree nodes.

Table 3
Decision rules derived from the obtained decision tree.

	Node	Decision rules
1	5	If TEMP is high and HLC > 3.89 then EUI is HIGH
2	6	If TEMP is low and HEAT is electric then EUI is HIGH
3	9	If TEMP is high and HLC ≤ 3.89 and ELA > 4.41 then EUI is LOW
4	10	If TEMP is low and HEAT is non-electric and NUM ≤ 2 then EUI is LOW
5	12	If TEMP is high and HLC ≤ 3.89 and ELA ≤ 4.41 and HWS is electric then EUI is LOW
6	15	If TEMP is low and HEAT is non-electric and NUM > 2 and HOUS is apartment then EUI is HIGH
7	16	If TEMP is high and HLC ≤ 3.89 and ELA ≤ 4.41 and HWS is non-electric and KITC is electric then EUI is LOW
8	18	If TEMP is low and Heat is non-electric and Num > 2 and HOUS is detached and HLC ≤ 1.70 then EUI is LOW
9	19	If TEMP is low and Heat is non-electric and Num > 2 and HOUS is detached and HLC > 1.70 then EUI is HIGH
10	20	If TEMP is high and HLC ≤ 3.89 and ELA ≤ 4.41 and HWS is non-electric and KITC is non-electric and HLC ≤ 2.93 then EUI is LOW
11	21	If TEMP is high and HLC ≤ 3.89 and ELA ≤ 4.41 and HWS is non-electric and KITC is non-electric and HLC > 2.93 then EUI is HIGH

```
  a   b  < -- classified as
 35   1  |   a = 'LOW EUI'
  3  16  |   b = 'HIGH EUI'
```

In this matrix, the number of correctly classified records is given in the main diagonal, i.e. upper-left to lower-right diagonal; the others are incorrectly classified. Clearly, class "LOW EUI" was misclassified as "HIGH EUI" only one time and class "HIGH EUI" was misclassified as "LOW EUI" three times. Such information indicates that high EUI is more prone to be misclassified than low EUI. This may have occurred due to the fact that most of the data records are in LOW EUI so the tree is made more sensitive to this class. An even distribution between HIGH EUI class and LOW EUI class in database would possibly help obtain sufficient accuracy and sensitivity in the desired classes.

The major strength of decision tree lies in its interpretability and ease of use, particularly when decision rules are created. Based on a decision tree, decision rules can be easily generated by traversing a path from the root node to a leaf node. For example, a decision rule can be generated from node 1 to node 5 in above decision tree as follows: If TEMP is high and HLC ≤ 3.89 and ELA ≤ 4.41 and HWS is electric then EUI is LOW. Since each leaf node produces a decision rule, the complete set of decision rules, which is equivalent to the decision tree, can be derived after all the leaf nodes have been included. Accordingly, above decision tree is converted to a set of decision rules, as show in Table 3.

4.2. Evaluation of the decision tree

As mentioned previously, the decision tree accuracy should be evaluated to estimate how accurately it can predict building EUI levels before applying it to new residential buildings. Accordingly, the obtained decision tree was applied to the test dataset and the results are given in Table 4.

Table 5 shows that among 12 data records included in the testing set eleven records, accounting for 92%, are correctly classified. Given that the size of testing set is relatively small and only one record is misclassified, this accuracy is basically acceptable. At the same time, WEKA analysis report also provides confidence level for the classification of each data record. The confidence level determines how likely the test data record falls into that class and, it is equal to the ratio of the number of correctly classified data records to total record number in that class in the training set. It can be seen from Table 5 that generally the confidence level for the classification is higher than 80%, indicating that most of the prediction is reliable. Further, by using a pre-specified threshold, e.g. 80%, confidence level could improve estimated accuracy of classification. In particular, if the confidence level of a data record classification exceeds the threshold, this classification will be accepted; otherwise it will be refused. For example, if the threshold in this evaluation is set to be 80%, then all the records, except the record 2 that is misclassified, will be accepted. Similarly, the threshold is very useful when applying decision rules to the prediction of new data sets. In addition, the error rate between the actual EUI value and the reference EUI value are also given in this table for the reliability test of reference value. It can be seen that, among 11 correctly classified data records, 5 have an error rate lower than 5% while the other 6 have an error rate between 20% and 35%, which indicates that a higher concept hierarchy for building EUI need to be formed to improve the prediction performance of reference value. However, this is limited by the size of database in this study.

4.3. Utilization of decision tree

4.3.1. Using decision tree for prediction

Based on predictor variables, decision tree and decision rules can be utilized to predict target Variables Assume the EUI level of a new residential building in Japan must be predicted by using the decision tree in Fig. 6. The threshold of confidence level is set to be 85%. The typical building parameters are shown in Table 5.

Specifically, the building EUI level is predicted as follows:

Step 1: The root node, i.e. node 1 in this decision tree, is the starting point of prediction. From node 1, it can be seen the value of TEMP

Table 4
Results of decision tree accuracy evaluation.

	Actual level	Predicted level	Correct or incorrect	Confidence level	Actual EUI	Reference EUI	Error
1	HIGH	HIGH	Correct	100%	449	450	0.2%
2	LOW	HIGH	Incorrect	75%	258	624	141.9%
3	HIGH	HIGH	Correct	100%	581	584	0.5%
4	LOW	LOW	Correct	100%	327	322	1.5%
5	HIGH	HIGH	Correct	100%	707	552	22.0%
6	LOW	LOW	Correct	81.80%	303	316	4.3%
7	LOW	LOW	Correct	81.80%	238	316	32.8%
8	LOW	LOW	Correct	88.90%	258	315	22.1%
9	HIGH	HIGH	Correct	100%	507	488	3.7%
10	HIGH	HIGH	Correct	100%	495	601	21.4%
11	LOW	LOW	Correct	81.80%	427	316	26.0%
12	HIGH	HIGH	Correct	100%	458	601	31.2%

Table 5
Building parameters for the prediction of building EUI levels.

Number	Variable	Attribute value	Unit
1	TEMP	High	
2	HOUS	Detached house	
3	CONS	Wood	
4	NUM	4	
5	AREA	100	m^2
6	HLC	2	W/m^2K
7	ELA	3	cm^2/m^2
8	HEAT	Electricity	
9	HWS	Non-electricity	
10	KITC	Gas	

should be first examined. Since TEMP is high, the node 1 test *TEMP is high* is satisfied, then go to node 2;
Step 2: examine the value of HLC. Since HLC = 2, the node 2 test $HLC \leq 3.89$ is satisfied, then go to node 4;
Step 3: examine the value of ELA. Since ELA = 3, the node 4 test $ELA \leq 4.41$ is satisfied, then go to node 8;
Step 4: examine the value of HWS. Since HWS is non-electric, the node 8 test *HWS is electric* is not satisfied, then go to node 13;
Step 5: examine the value of KITC. Since KITC is gas, the node 13 test *KITC is electric* is not satisfied, then go to node 17;
Step 6: examine the value of HLC. Since HLC = 2, the node 17 test $HLC \leq 2.93$ is satisfied, then go to node 20;
Step 7: node 20 is a leaf node. As a result, the decision tree in Fig. 6 predicts that the EUI level of the residential building is LOW. In this node, the correctly classified data records account for 89% and thus the confidence level of prediction is 89% that is larger than the predetermined threshold (85%). Therefore, the prediction is accepted. Furthermore, the value of correctly classified records in this node ranges from 242 MJ/m^2 to 389 MJ/m^2 and the average value is calculated at 315 MJ/m^2. These values can be used as reference values for the prediction, as mentioned previously.

4.3.2. Model interpretation and useful information extraction

Useful information can be extracted from the decision tree based model so as to help understand energy consumption patterns and optimize a building design plan. For example, various parameters are automatically selected as predictor variables by the decision tree algorithm for the classification of EUI levels. These parameters are used to split the nodes of the decision tree and their degrees of closeness to the root node indicate the strength of the influence and the number of records impacted. Therefore, by examining the decision tree nodes, the significant factors, as well as their ranks, that determine the building energy demand profiles can be identified. In particular, the variable importance of this decision tree model can be analyzed as follows: first, the root node, i.e. TEMP, indicates that outside air temperature is the most important determinant of energy demand among all these factors. Then, for clarity, the significant factors for the high temperature districts (i.e. Hokuriku, Kanto, Kansai and Kyusyu) and low temperature districts (i.e. Hokkaido and Tohoku) are identified separately and summarized in Table 6.

Clearly, four significant factors are identified for each district and the only parameter found to be significant for the both districts is heat loss coefficient. This implies that the significance of these factors, except building heat loss coefficient, is dependent on outside air temperature. Moreover, among the three household appliance energy source parameters, space heating plays a role in low temperature districts while hot water supply and kitchen are significant in high temperature districts. Note that floor area and construction types do not appear in the decision tree. This is reasonable since the target variable, i.e. EUI level, is a measure of annual total energy normalized for floor area and building heat loss coefficient embodies the effect of construction type. At the same time, these significant factors are ranked in terms of the degree of closeness to the root node. It can be found that heat loss coefficient and space heating mode rank the first in the two districts respectively, and thus deserve extra attention when designing energy efficient buildings.

Table 6
Summary of significant factors.

Potential factors	High temperature districts		Low temperature districts	
	Significant factors	Rank	Significant factors	Rank
House type			√	3
Number of occupants			√	2
Floor area				
Heat loss coefficient	√	1	√	4
Equivalent leakage area	√	2		
Construction type				
Space heating mode			√	1
Hot water supply mode	√	3		
Kitchen energy mode	√	4		

The decision tree can provide the combination of significant factors as well as the threshold values that will lead to high building energy performance. Based on such combination and threshold values, some hidden yet useful information can also be extracted to help understand building energy consumption patterns. For example, it can be seen that, in high temperature districts, a higher building heat loss coefficient than 3.89 W/m^2K will normally cause a high EUI. Meanwhile, for a residential building with heat loss coefficient lower than 3.89 W/m^2K, a high equivalent leakage area (>4.41 cm^2/m^2) will benefit energy conservation. This seems perhaps unreasonable and one possible explanation is that the high temperature districts locate in moderate climate and have a moderate outside air temperature range. Accordingly, in summer infiltration can serve as cooling source to remove the excess heat generated indoor, thereby reducing overall energy consumption. This indicates that a rational combination of heat loss coefficient and equivalent leakage area of residential buildings in high temperature districts is important to improve building energy performance. Also, a further study on the range selection of equivalent leakage area may provide deeper insights into its impact on building energy demand. Additionally, from the nodes 8 and 13 in Fig. 6, it can be observed that the change of the energy source of hot water supply and kitchen will bring about a substantial increase or decrease in EUI. Clearly electrical water heaters, instead of non-electric water heaters such as natural gas heaters, should be used to save energy. Moreover, electrical water heaters can take full advantage of cheap nighttime electricity and thus help users save money spent on energy.

The EUI values in the node 8 are plotted in Fig. 8 in order to make a comparison between buildings with electric HWS and buildings with non-electric HWS. The two significant factors with higher ranks than HWS, i.e. HLC and ELA, are also taken into consideration (HLC at abscissa, ELA at ordinate). The abscissa–ordinate plane is divided into various grids so that EUI values can be compared based on similar HLC and ELA values, thereby removing the impact of these two factors. It is apparent from Fig. 8 that, in a same grid or adjacent grids, red points, which denote EUI values with non-electric HWS, are generally higher than blue points, which denote EUI values with electric HWS. This is in accordance with the above conclusion drawn from the decision tree.

With regard to kitchen energy source, electrical appliances, however, tend to consume more energy than the appliances using natural gas. This may have occurred since the power of many kitchen electrical appliances, such as rice cooker, is comparatively

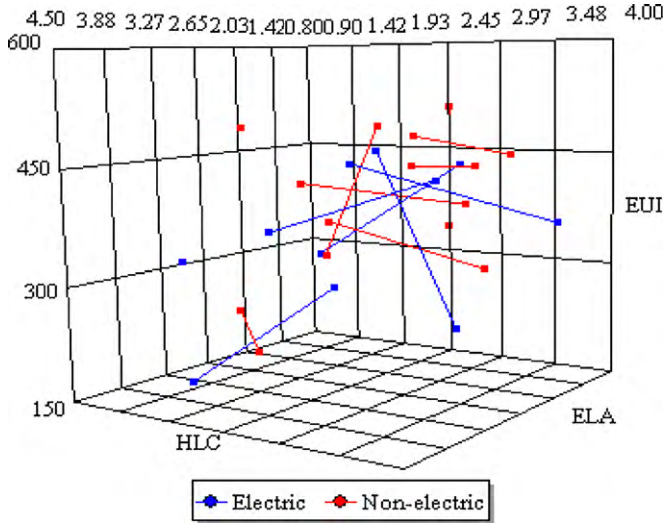

Fig. 8. Comparison of EUI between electric HWS and non-electric HWS.

high and the use of these appliances is routine. Further, compared to hot water supply energy source, kitchen energy source has a smaller contribution to building energy demand and even though non-electric appliances is adopted in kitchen, an extra requirement on heat loss coefficient ($\leq$2.93 W/m^{2}K) still need to be met in order to achieve low EUI levels.

In low temperature districts, from an energy saving point of view, building owners and designers should give a prior consideration to space heating energy source that plays a significant role in influencing EUI. The node 3 in Fig. 6 shows that non-electric fuel, particularly kerosene and natural gas, should be used as primary source of residential space heating since the use of electric space heating tends to bring about a high EUI. This may be partly ascribed to the high efficiency of non-electric space heating devices such as kerosene space heaters. Moreover, non-electric heating devices are more applicable than electric space heaters, such as air conditioners, in real life due to the high electricity rate in Japan. Similar to Fig. 8, EUI values in the node 3, together with EUI values in low temperature districts in the test dataset, are plotted in Fig. 9. HLC and NUM are used as abscissa and ordinate. The red and blue points represent EUI values with electric and non-electric space heating respectively. It can be observed that red points are generally higher than blue points, which is in accordance with above conclusion.

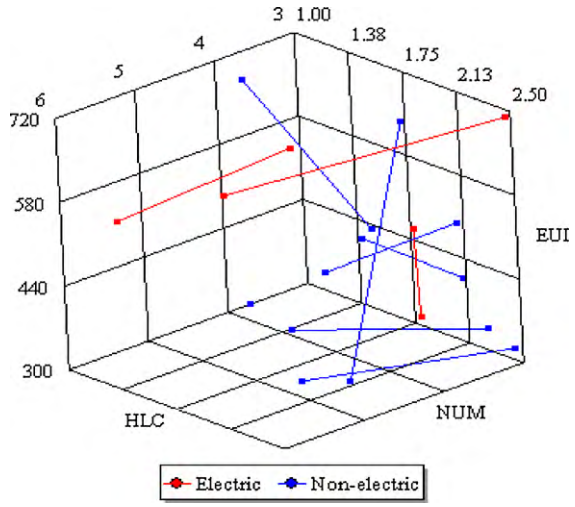

Fig. 9. Comparison of EUI between electric HEAT and non-electric HEAT.

Family size, i.e. the number of occupants, is another important determinant of EUI in low temperature districts. As can be seen, families with more than two occupants will have significantly higher EUI than those with two occupants. This may have occurred since a larger family size will cause more complicated occupant behavior patterns thereby resulting in an increase in EUI. With regard to house type, it can be seen that detached houses with low heat loss coefficients ($\leq$1.70 W/m^{2}K) tend to have a better energy performance than apartments, which can occur for at least two reasons. First, a small HLC contributes greatly to reduce energy consumption on space heating and cooling; second, detached houses normally have larger areas than apartments while both of them have approximately same family size, which also lowers EUI values.

Such information can help building designers and owners make intelligent decisions to improve building energy performance and reduce building energy consumption. For example, based on above information, architects and building designers can identify the parameter that deserves more attention as well as its value range at the early design stage. Also, they can perform a fast performance estimation of newly constructed residential buildings. Moreover, building owners will easily determine which energy source should be used for space heating, hot water supply, and kitchen to save energy. It should be mentioned that heat loss coefficient and equivalent leakage area cannot be determined directly by architects and building designers. However, their value can be adjusted through some indirect measures such as improving construction material and building air tightness.

5. Conclusions

In this paper, a decision tree method is proposed for building energy demand modeling. This method is applied to Japanese residential buildings for predicting and classifying building EUI levels and its basic steps, such as the generation of decision tree based on training data and the evaluation of decision tree based on test data are presented. The results have demonstrated that the use of decision tree method can classify and predict building energy demand levels accurately (93% for training data and 92% for test data), identify and rank significant factors of building EUI levels automatically, and provide the combination of significant factors as well as the threshold values that will lead to high building energy performance. Such method along with derived information could benefit building owners and designers greatly and one crucial benefit is improving building energy performance and reducing energy consumption and the money spent on energy. Although the decision tree method is mainly employed to predict categorical variables (the number of the predetermined target intervals depends on the size of database while too many intervals may result in errors in classification) and reference value (i.e. average value of EUI in each class in this study) instead of the precise value of target variables, as a modeling technique, the utilization of decision tree method is very simple and its result can be interpreted more easily compared to other widely used modeling techniques, such as regression method and ANN method.

The application of decision tree method to Japanese residential buildings in this paper has clearly demonstrated that this method is feasible, having many advantages over other modeling techniques. However, further study still need to be carried out to provide deeper insights into the utilization of this method to modeling building energy demand. The main focus of future research should be placed on selecting appropriate interval number and reference value of target variables without reducing estimation accuracy, since these measures will provide more precise and valuable information to users. In addition, more case studies in different sectors, such as commercial buildings and office buildings, should be conducted to further benefit energy conservation and policy formulation.

Acknowledgements

The authors would like to express their gratitude to the Public Works and Government Services Canada, and Concordia University for the financial support.

References

[1] L. Pérez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy and Buildings 40 (3) (2008) 394–398.
[2] W.G. Cai, Y. Wu, Y. Zhong, H. Ren, China building energy consumption: situation, challenges and corresponding measures, Energy Policy 37 (6) (2009) 2054–2059.
[3] T. Catalina, J. Virgone, E. Blanco, Development and validation of regression models to predict monthly heating demand for residential buildings, Energy and Buildings 40 (10) (2008) 1825–1832.
[4] C. Ghiaus, Experimental estimation of building energy performance by robust regression, Energy and Buildings 38 (6) (2006) 582–587.
[5] L. Zhou, F. Haghighat, Optimization of ventilation systems in office environment. Part I. methodology, Building and Environment 44 (2008) 651–656.
[6] L. Magnier, F. Haghighat, Multiobjective optimization of building design using genetic algorithm and artificial neural network, Building and Environment 45 (2010) 739–746.
[7] J. Zhang, F. Haghighat, Development of artificial neural network based heat convection for thermal simulation of large rectangular cross-sectional area earth-to-earth heat exchanges, Energy and Buildings 42 (4) (2010) 435–440.
[8] Y.P. Zhou, J.Y. Wu, R.Z. Wang, S. Shiochi, Y.M. Li, Simulation and experimental validation of the variable-refrigerant-volume (VRV) air-conditioning system in EnergyPlus, Energy and Buildings 40 (6) (2008) 1041–1047.
[9] F.F. Al-ajmi, V.I. Hanby, Simulation of energy consumption for Kuwaiti domestic buildings, Energy and Buildings 40 (6) (2008) 1101–1109.
[10] L. Wehenkel, M. Pavella, Decision tree approach to power systems security assessment, International Journal of Electrical Power & Energy Systems 15 (1) (1993) 13–36.
[11] C.-Y. Fan, P.-C. Chang, J.-J. Lin, J. C. Hsieh. A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Applied Soft Computing, in press, Corrected Proof.
[12] K.-Y. Tung, I.-C. Huang, S.-L. Chen, C.-T. Shih, Mining the Generation Xers' job attitudes by artificial neural network and decision tree—empirical evidence in Taiwan, Expert Systems with Applications 29 (4) (2005) 783–794.
[13] G.K.F. Tso, K.K.W. Yau, Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks, Energy 32 (9) (2007) 1761–1768.
[14] J.R. Quinlan, Induction of decision trees, Machine Learning (1986).
[15] J. Han, M. Kamber, Data mining concepts and techniques, Elsevier Inc., San Francisco, 2006.
[16] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth Inc., 1984.
[17] J.R. Quinlan, C4. 5 Programs for Machine Learning, Morgan Kaufmann, San Mateo, 1993.
[18] C.E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–623.
[19] S. Murakami, S-i. Akabayashi, T. Inoue, H. Yoshino, K-i. Hasegawa, K. Yuasa, T. Ikaga, Energy Consumption for Residential Buildings in Japan, Architectural Institute of Japan, Maruzen Corp., 2006, http://www.jma.go.jp/jma/indexe.html.