JOURNAL OF CIVIL ENGINEERING AND MANAGEMENT
ISSN 1392-3730 / eISSN 1822-3605
2015 Volume 21(6): 748–760
doi:10.3846/13923730.2014.893908

PREDICTION OF GOVERNMENT-OWNED BUILDING ENERGY CONSUMPTION BASED ON AN RRELIEFF AND SUPPORT VECTOR MACHINE MODEL

Hyojoo SON[a], Changmin KIM[a], Changwan KIM[a], Youngcheol KANG[b]

[a]*Department of Architectural Engineering, Chung-Ang University, 156-756 Seoul, Korea*
[b]*Department of Global Construction Management, The University of Seoul, Liberal Arts Building 5224, 163 Siripdaero, Dongdaemun-gu, 130-743 Seoul, Korea*

Received 24 Sep 2012; accepted 15 Jan 2013

Abstract. Accurate prediction of the energy consumption of government-owned buildings in the design phase is vital for government agencies, as it enables formulation of the early phases of development of such buildings with a view to reducing their environmental impact. The aim of this study was to identify the variables that are associated with energy consumption in government-owned buildings and to propose a predictive model based on those variables. The proposed approach selects relevant variables using the RReliefF variable selection algorithm. The support vector machine (SVM) method is used to develop a model of energy consumption based on the identified variables. The proposed approach was analyzed and validated on data for 175 government-owned buildings derived from the 2003 Commercial Building Energy Consumption Survey (CBECS) database. The experimental results revealed that the proposed model is able to predict the energy consumption of government-owned buildings in the design phase with a reasonable level of accuracy. The proposed model could be beneficial in guiding government agencies in developing early strategies and proactively reducing the environmental impact of a building, thereby achieving a high degree of sustainability of buildings constructed for government agencies.

Keywords: energy consumption prediction, government-owned building, RReliefF variable selection, support vector machine model, sustainable development.

Introduction

The building sector is one of the biggest contributors to worldwide energy consumption and environmental pollution. For example, the building sector is responsible for 40% of energy consumption in the European Union (EU) (International Energy Agency 2010) and more than 40% in the United States (U.S.) (U.S. Department of Energy 2012). Aside from playing a major role in energy consumption, the building sector is among the sectors most responsible for environmental degradation, amounting to 36% of carbon dioxide (CO_2) emissions in the EU (European Commission 2012) and 39% of CO_2 emissions in the U.S. (U.S. Department of Energy 2012). Moreover, energy consumption in the building sector is expected to grow, as the need for new buildings continues to be spurred by the growth of the world economy and development.

The most cost-effective solution for reducing building energy consumption is to design energy-efficient buildings. Incompetent building design (in terms of energy efficiency) could have a significant impact on an owner's financial risks as well as on the sustainability of development. Specifically, implementation of energy-efficient design in new government buildings has become a top priority for government agencies. For example, U.S. federal agencies are required to reduce their energy intensity by 30% by the end of 2015 relative to 2003 levels, under Executive Order 13423, issued on January 24, 2007 (U.S. National Archives and Records Administration 2007). To implement energy-efficient building design, project stakeholders need to understand the effects of design decisions on the energy performance of the building to be built. Further, to account for the effects of individual design features on energy consumption, an accurate energy consumption model is needed.

Energy consumption models have been constructed using statistical regression and machine learning methods. Regression analysis is the most widely used technique for the modeling of relationships between design features representing important characteristics of buildings and building energy consumption (see, for example, Sharp 1996; Chung *et al.* 2006; Lee 2008; Chung, Hui 2009). However, the disadvantage of using regression analysis with the applications in the real world, e.g. modeling of building energy consumption,

Corresponding author: Changwan Kim
E-mail: *changwan@cau.ac.kr*

www.tandfonline.com/tcem

Taylor & Francis
Taylor & Francis Group

is that this method assumes linearity in the relationship between the dependent and independent variables and normality in the error distribution, which may not be valid for a given data set (Kumar, Bhattacharya 2006; Chen 2011; Huang 2011; Li, Sun 2011).

Recently, machine learning methods have been applied to obtain reliable predictive models of building energy consumption. The artificial neural network (ANN) is the main computational model that has been used for this purpose. Yalcintas (2006) employed an ANN based on the Levenberg–Marquardt back-propagation algorithm to predict annual electricity consumption per square foot using data from 63 buildings in Hawaii. Ten input variables were used in the model: three plug-load-related variables (*computers, fume hoods,* and *other equipment*), four lighting-related variables (*lighting hours, floor percentage lighted, internal lighting type,* and *external lighting type*), and three HVAC-related variables (*HVAC hours, floor percentage air conditioned,* and *HVAC equipment type*). The correlation coefficient for the predicted and actual energy use was 0.86. However, no comprehensive guidance such as mean absolute error (MAE), root mean squared error (RMSE), or mean absolute percentage error (MAPE) was available to enable evaluation of the model in terms of the accuracy of the predicted data relative to that of the actual data.

Yalcintas and Ozturk (2007) used an ANN based on the Levenberg–Marquardt back-propagation algorithm with multiple linear regression (MLR) to predict annual electricity consumption (in kilowatt hours) per square meter using data from the Commercial Building Energy Consumption Survey (CBECS) Database. ANN and MLR models were constructed using eight input variables for each of nine census divisions. Sample sizes for each census division varied from 57 to 221. Yalcintas and Ozturk (2007) used only eight of the more than 300 variables available in the CBECS database, but they gave no explanation of their selection of those variables. The input variables were *building-operation hours, age category, building area per worker category, building area per computer, cooling percentage category, lighting percentage category, cooling degree days,* and *number of floors category*. Comparisons were made on the basis of the mean squared error (MSE) of the estimators. The MSE ranged from 9.60 to 15.25 for the ANN, and from 10.24 to 40.43 for the MLR model. It was found that the ANN produced better predictions than the MLR model. The main advantage of ANNs over regression models stems from their ability to model nonlinear relationships without needing to make assumptions in the data generating process (Hornik *et al.* 1990). However, construction of an ANN model has its drawbacks, as it requires a large quantity of training data in order to be trained properly, calls for numerous controlling parameters, presents difficulties in regard to obtaining a stable solution, and has a considerable likelihood of over-fitting.

The decision tree method has also been used to predict building energy consumption. Yu *et al.* (2010) reported on the development of a predictive model of building energy demand based on the decision tree method. This method is able to classify and predict categorical variables. It has a competitive advantage over other widely used modeling techniques, such as the regression method and the ANN method, in that it can generate accurate predictive models with interpretable flowchart-like tree structures that enable users to quickly extract useful information. To demonstrate its applicability, Yu *et al.* (2010) applied the method to the estimation of residential building energy performance indexes by modeling building energy use intensity (EUI) levels. The results demonstrated that the decision tree method is able to classify and predict building energy demand levels accurately (93% for training data and 92% for test data) and that it can automatically identify and rank the variables that have a significant effect on building EUI.

Previous studies of energy consumption focused mainly on methods for predicting the energy consumption of commercial or private buildings. To our knowledge, there has been no research that addresses the prediction of energy consumption of government-owned buildings. Hence, there is relatively little understanding of the variables that contribute to the prediction of energy consumption in government-owned buildings. Because of the excellent performance of support vector machines (SVMs) in general, they been used in a wide variety of applications. The theory of an SVM is founded on the structural risk minimization (SRM) principle (Vapnik 1995), which has exhibited better performance than the traditional empirical risk minimization (ERM) principle employed by conventional neural networks (Schölkopf *et al.* 1999; Wang *et al.* 2003). SRM minimizes an upper bound on the generalization error and allows an SVM to generalize better than an ANN.

The aims of the current study were to identify variables associated with energy consumption of government-owned buildings by using the RReliefF variable selection algorithm, and to propose a predictive model for the energy consumption of government-owned buildings, by using an SVM model based on the identified variables. To evaluate the prediction performance of the SVM model, we performed a comprehensive comparison of the prediction performance of the SVM model versus that of the ANN model, the DT model, and the MLR model. A data set on government-owned office buildings was taken from the 2003 CBECS database and used for training and testing experiments. In Section 1 we present some material on the RReliefF variable selection algorithm and the SVM model, and in Section 2 we describe the data source and the experimental settings. In Section 3 we present an explanation of our implementation of the RReliefF variable selection algorithm and our development of the SVM, ANN, DT, and MLR models, as well as a discussion and analysis of the experimental results. The final section contains our conclusions and suggests directions for future research.

1. Theoretical background

1.1. RReliefF

From a machine learning perspective, the input variables used in constructing a predictive model are not all expected to be of equal importance or quality (Yang *et al.* 2008). In fact, the inclusion of irrelevant or redundant variables could reduce the performance of a machine learning algorithm (Robnik-Šikonja, Kononenko 2003). Although expert knowledge of the application domain can still be used as a guide to identify relevant variables, there is relatively little understanding of which variables have a significant effect on the prediction of energy consumption in government-owned buildings. One possible solution to this problem is to extract a number of candidate variables characterizing building energy consumption prediction and then implement a variable selection algorithm to identify the relevant variables (Molina *et al.* 2002). One of the strongest benefits of variable selection algorithms developed recently is that they can improve the prediction accuracy while reducing the dimensionality of the input space by searching and finding the optimal variable subset.

In this study, the RReliefF algorithm, which was proposed by Robnik-Šikonja and Kononenko (1997), was used to select the optimal variable subset. RReliefF considers contextual information and effectively and correctly takes into account interdependencies between variables (Robnik-Šikonja, Kononenko 2003). Because of the variable interdependency feature, RReliefF is better than variable selection algorithms based solely on statistical measures such as the correlation coefficient, information gain, and the signal-to-noise ratio (Robnik-Šikonja, Kononenko 2003; Yang *et al.* 2008). In addition, its capabilities, with respect to noise, are robust because it searches and selects the nearest neighbors to determine the importance of each variable (Pernek *et al.* 2012).

RReliefF is an adaptation of ReliefF (Kononenko 1994) for solving a regression problem. ReliefF is, in turn, an extension of Relief (Kira, Rendell 1992) for solving multi-class (more than two-class) classification problems (Kononenko *et al.* 1996; Pham *et al.* 2009; Kandaswamy *et al.* 2011; Han, Yu 2012). The base algorithm for Relief was limited to binary (two-class) classification problems. The main idea behind the Relief algorithm is that high-quality and highly relevant variables should distinguish between instances from different classes and should have similar values for instances from the same class. Specifically, Relief ranks the variables based on quantification of how well they satisfy those two conditions.

Relief evaluates the variables one by one and assigns a real number to each variable to indicate its importance. Relief randomly selects an instance *R* from the data set and finds the nearest neighbor *H* from the same class (the nearest hit) and the nearest neighbor *M* from the other class (the nearest miss). Then it updates the score for each variable by comparing the value of that variable in *R* with its values in *H* and *M*. If *R* and *H* have different values of some variable *f*, this means that two instances from the same class can be falsely separated by *f* (not desirable), so *f*'s score is decreased. If *R* and *M* have different values of *f*, this means that two instances from different classes can be correctly separated by *f* (desirable), so *f*'s score is increased. The process is repeated for a number of randomly selected instances from the data set.

When solving the regression problem, nearest hits and nearest misses cannot be used, because, in general, continuous (not just discrete) variables can be used in the construction of predictive models. Thus instead of requiring knowledge of whether two instances belong to the same or a different class, RReliefF attempts to solve this problem by introducing a score which indicates that the predicted values of a given independent variable in the two instances are different. The score is based on the relative distance between these values. The scores that are calculated by RReliefF lie in the range [–1, 1]. At the extremes, a score of +1 signifies that different values of a variable imply different values of the dependent variable for nearby instances, while a score of –1 signifies that different values of a variable imply equal values of the dependent variable for nearby instances. In other words, the larger the value of the RReliefF score, the greater the influence of that variable on the regression. An RReliefF value greater than zero indicates that that variable is likely to distinguish between nearby instances to a useful extent. In this study, the significance threshold value was set at 0.01. This threshold value was chosen according to previous studies by Robnik-Šikonja, Kononenko (2003) and Srisawat, Kijsirikul (2009) to obtain the limited number of variables needed to separate the relevant variables from the irrelevant ones.

1.2. Support Vector Machine (SVM)

The roots of the SVM date back to the discrimination work of Vapnik and Lerner (1963). The general nonlinear version of the SVM for classification problems was introduced much more recently (Vapnik 1995), but it wasn't until even later that the theory was extended to the solution of a nonlinear regression using the nonlinear version of the SVM framework (Suykens, Vandewalle 1998). A simple description of the support vector machine algorithm for regression is provided in what follows.

Consider a data set of the form $\left(x_i, y_i\right)_{i=1}^{m}$, where the inputs are *n*-dimensional input vectors x_i with real-valued components, the outputs y_i are the corresponding values of a real-valued dependent variable, and *m* is the total number of points in the data set. The objective of the regression analysis is to find a regression function *f*(*x*) that accurately predicts the outputs. In an SVM, solving a nonlinear regression problem requires that the input vectors first be nonlinearly mapped into a high-dimensional input space and then linearly correlated with the outputs. The SVM formalism uses the following linear estimation function:

$$f(x) = w \cdot \varphi(x) + b, \tag{1}$$

where: *x* is an input vector; $w \cdot \varphi(x)$ is the weight vector; φ is the nonlinear map; w is the dot product of *w* and φ(*x*); and *b* is a constant.

In the SVM formulation, the ε-insensitive loss function L_ε is used as a cost function:

$$L_\varepsilon(f(x), y) = \begin{Bmatrix} 0 & \text{if} \left|f(x) - y\right| < \varepsilon \\ \left|f(x) - y\right| - \varepsilon & \text{otherwise} \end{Bmatrix}, \quad (2)$$

where ε is a precision parameter that represents the radius of the tube located around the regression function *f*(*x*).

The weight vector *w* and the constant *b* in Eqn (1) can be estimated by minimizing the following regularized risk function:

$$R(C) = C\frac{1}{m}\sum_{i=1}^{m} L_\varepsilon(f(x_i), y_i) + \frac{1}{2}\|w\|^2, \quad (3)$$

where $\frac{1}{2}\|w\|^2$ is a regularization term that controls the trade-off between the complexity and the approximation accuracy of the regression model and ensures that the model possesses improved generalized performance. *C* is a regularization constant that accounts for the trade-off between the empirical risk and the regularization term.

Two positive slack variables, ξ_i and ξ_i^*, can be used to measure the deviation of $y_i - f(x_i)$ from the boundaries of the ε-insensitive zone. When these slack variables are used, Eqn (3) is transformed into the constrained form:

$$R(w, \xi_i^*) = \frac{1}{2}\|w\|^2 + C\sum\nolimits_{i=1}^{m}(\xi_i + \xi_i^*), \quad (4)$$

where the minimization is subject to the following constraints:

$$(w \cdot \phi(x_i)) + b - y_i \geq -(\varepsilon + \xi_i);$$

$$(w \cdot \phi(x_i)) + b - y_i \leq \varepsilon + \xi_i^*;$$

$$\xi_i, \xi_i^* \geq 0.$$

Use of Lagrange multipliers and Karush–Kuhn–Tucker conditions in Eqn (4) yields the dual Lagrangian form:

$$L_d(\alpha, \alpha^*) = -\varepsilon\sum_{i=1}^{m}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{m}(\alpha_i^* - \alpha_i) y_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j), \quad (5)$$

where the minimization is subject to the following constraints:

$$\sum_{i=1}^{m}(\alpha_i^* - \alpha_i) = 0;$$

$$0 \leq \alpha_i \leq C;$$

$$0 \leq \alpha_i^* \leq C.$$

K is the kernel function and the Lagrange multipliers in Eqn (5), α_i and α_i^*, must satisfy the condition $\alpha_i \alpha_i^* = 0$. The optimal desired weight vector of the regression hyperplane is $v^* = \sum\nolimits_{i=1}^{m}(\alpha_i - \alpha_i^*)K(x, x_i)$. It can be shown that the general form of an SVM-based regression function can be written as:

$$f(x, \alpha, \alpha^*) = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)K(x, x_i) + b. \quad (6)$$

Some kernel functions render it easier to obtain the optimal solution in an SVM than others do. The three most frequently used kernel functions are polynomial, sigmoid, and radial basis functions (RBFs), all of which are frequently used because, unlike linear kernel functions, they can classify multi-dimensional data. Also, RBFs have fewer parameters to set than polynomial kernels do, but their overall performance is similar to that of other kernel functions. Therefore, in this study, an RBF was used as the kernel function in the SVM to obtain the optimal solution. A radial basis kernel function is of the form:

$$K(x_i, x_j) = \exp(-\gamma\left\|x_i - x_j\right\|^2), \quad (7)$$

where γ is the so-called kernel parameter.

2. Data set and pre-processing

In this study, we used data on government-owned buildings from the 2003 Commercial Buildings Energy Consumption Survey (CBECS) database to evaluate the predictive performance of an energy consumption model. The 2003 CBECS contains data on energy consumption, energy expenditure, and extensive energy-related building characteristics for 1,057 government-owned buildings that are categorized according to ownership as either local, state, or federal government. The only buildings we included in our analysis are those with a total gross floor area of at least 2,000 square meters that are used for 12 months per year and more than 40 hours per week. This narrowed the study to 526 government-owned buildings, and these were used to estimate building energy consumption by modeling building energy use intensity (EUI) levels.

After an in-depth examination of all the independent variables for which data are provided in the 2003 CBECS, we chose the ones that could possibly be of relevance to this study. There were 26 such variables in that subset, and they were grouped into the following categories: general building information and energy end uses; building activities and special measures of size; heating and cooling equipment and conservation features; water heating, refrigeration, office equipment, and special uses of space; and lighting percents, equipment, and conservation features. The input variables used in our energy analysis are listed in Table 1. However, we adjusted the value of the "number of floors" input variable. In the 2003 CBECS database, the number of floors for buildings with more than 14 floors is given as either 991 or 992, where 991 indicates that a building has 15–25 floors, and 992 indicates that it has more than 25 floors. In this study, we replaced 991 with 19, and 992 with 32, to better represent the number of floors. The output variable O used in this analysis was electricity consumption per square meter per hour of operation:

$$O\ (\text{Wh/m}^2\ \text{per hour}) = \frac{A\ (\text{Wh})}{T\ (\text{hr}) \times S\ (\text{m}^2)}, \qquad (8)$$

where: A is the total annual electricity consumption (in watt hours, Wh); T is the total yearly operating time (in hr); and S is the total floor space (in m^2).

Buildings with missing values were omitted from the data set, resulting in a total of only 181 buildings to be included in the data analysis. To determine outliers in a distribution, we used the 1.5 × IQR criterion, which is the standard rule of thumb used in statistics for identifying suspected outliers (Moore, McCabe 1999). IQR is the interquartile range, the difference between the first quartile (Q1) and the third quartile (Q3). Items of value d were considered as a suspected outlier if $d > Q3 + 1.5 \times IQR$ or $d < Q1 - 1.5 \times IQR$ and were removed from the data set. After a data-cleaning step, we had a set of data for 175 government-owned buildings.

In an SVM model, each data instance is represented as a vector of real numbers. Hence, if there are categorical variables, they have to be converted into numeric data. In order to represent an m-category variable, m numbers are used. Only one of the m numbers is equal to 1, and the others are 0. For example, if the "window glass type" variable has four categories (such as single-layer glass, multi-layer glass, combination of single-layer glass and multi-layer glass, and no windows), then the value of that variable is represented as either (0, 0, 0, 1), (0, 0, 1, 0), (0, 1, 0, 0), or (1, 0, 0, 0).

Scaling is very important. The main advantage of scaling is to avoid having variables with large numeric ranges dominating those with smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. Because large values of variables could cause numerical problems. Therefore, in this study

Table 1. Description of the input variables used in this study

No.	Category	Variable name	Variable description
1	File 1[a]	CLIMATE8	Climate zone (30-year average)
2	File 1	WLCNS8	Wall construction material
3	File 1	RFCNS8	Roof construction material
4	File 1	GLSSPC8	Percent exterior glass
5	File 1	EQGLSS8	Equal glass on all sides
6	File 1	BLDSHP8	Building shape
7	File 1	NFLOOR8	Number of floors
8	File 1	OWNOCC8	Owner occupies space
9	File 2[b]	PBAPLUS8	More specific building activity
10	File 3[c]	HEATP8	Percent heated
11	File 3	MAINHT8	Main heating equipment
12	File 3	COOLP8	Percent cooled
13	File 3	MAINCL8	Main cooling equipment
14	File 3	VAV8	VAV system
15	File 3	ECN8	Economizer cycle
16	File 3	EMCS8	Energy management and control system
17	File 4[d]	PCNUM8	Number of computers
18	File 7[e]	LTOHRP8	Percent lighted when open
19	File 7	LTNHRP8	Percent lighted when closed
20	File 7	WINTYP8	Window glass type
21	File 7	TINT8	Tinted window glass
22	File 7	REFL8	Reflective window glass
23	File 7	AWN8	External overhangs or awnings
24	File 7	SKYLT8	Skylights/atriums designed for lighting
25	File 7	AUTOLT8	Auto controls or sensors on lighting
26	File 7	DAYLTP8	Percent daylight

[a] File 1: general building information and energy end uses;
[b] File 2: building activities and special measures of size;
[c] File 3: heating and cooling equipment and conservation features;
[d] File 4: water heating, refrigeration, and office equipment (and special space uses);
[e] File 7: lighting percents, equipment, and conservation features.

we used linear scaling and independently normalized each variable to the range [0, 1], which not only ensures that the variables with large values do not overwhelm those with smaller values but also helps to reduce prediction errors.

3. Methodology

3.1. Analysis steps

This study was conducted according to the procedure outlined in Figure 1. In Step 1, the RReliefF algorithm was applied to select the most relevant variables from the 26 input variables. The selected variables were those with a score of at least 0.01. As a result, the number of multi-dimensional input variables was reduced. In Step 2, a predictive model was constructed by training the SVM model with the variables selected in the first step. In Step 3, the ANN, DT, and MLR models were trained with those same variables, so that their performance could be directly compared to that of the proposed SVM model. In Step 4, the prediction performance of the four models was compared.

3.2. Variable selection using RReliefF

The first step of the procedure consisted of selecting the most relevant variables from the set of 26 candidate variables, using the RReliefF algorithm, which ranked the 26 variables by level of importance. We used all 175 instances (all 175 buildings in our final data set) in this step, to increase the reliability of the scores assigned to the variables.

Only seven of the 26 candidate variables had had a score of at least 0.01, so they were the variables selected for this study. Those variables and their scores are given in Table 2, where they are listed in descending order of their scores. As shown there, energy consumption in government-owned buildings can best be explained by variables such as *main heating equipment, climate zone (30-year average), more specific building activity, roof construction material, window glass type, main cooling equipment,* and *auto controls or sensors on lighting*.

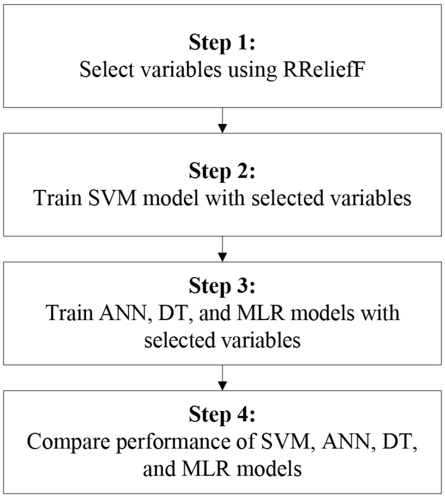
Step 1:
Select variables using RReliefF
Step 2:
Train SVM model with selected variables
Step 3:
Train ANN, DT, and MLR models with selected variables
Step 4:
Compare performance of SVM, ANN, DT, and MLR models

Fig. 1. Outline of procedure

3.3. SVM implementation

The seven variables obtained in the variable selection step were passed to the SVM predictive model for validation. A proper parameter setting can improve the prediction performance of an SVM. With an RBF kernel, there are two parameters to be determined in the SVM model: the regularization parameter C and the kernel parameter γ. Proper selection of (C, γ) could improve the prediction performance of the SVM. The "grid-search" approach proposed by Hsu *et al.* (2003) and Wang *et al.* (2005) is an alternative to finding the best C and γ when using an RBF kernel function.

In the grid-search approach, a common strategy is to separate the data set into two parts, one of which is considered unknown. The prediction accuracy obtained from the "unknown" set more precisely reflects the accuracy of predicting an independent data set. An improved version of this procedure is known as "cross-validation". The main advantage of cross-validation is to avoid the commonly occurring problem of overfitting (Kohavi 1995; Salzberg 1997). In k-fold cross validation, we first divide the training set into k subsets of equal size. Then each of the subsets is tested against the model that has been trained on the remaining k–1 subsets. Thus, each instance from the entire training set is predicted once. In this study, the grid search was carried out using 5-fold cross-validation.

The basic concept of a grid search is that various pairs of (C, γ) values are tried, and the one with the best cross-validation performance is picked. In the grid search used in this study, the pair (C, γ) with the lowest root mean squared error (RMSE) was chosen:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2}, \qquad (9)$$

where: the y_i are the actual values of the dependent variable; $f(x_i)$ are the corresponding values predicted by the model, and m is the number of points in the data set.

Table 2. RReliefF scores of the selected variables

No.	Variable name	Variable description	Score
1	MAINHT8	Main heating equipment	0.0848
2	CLIMATE8	Climate zone (30-year average)	0.0470
3	PBAPLUS8	More specific building activity	0.0450
4	RFCNS8	Roof construction material	0.0181
5	WINTYP8	Window glass type	0.0177
6	MAINCL8	Main cooling equipment	0.0138
7	AUTOLT8	Auto controls or sensors on lighting	0.0131

The grid search was performed as follows (Hsu *et al.* 2003): first, we selected a grid space with $C \in \{0.1,\ldots 50{,}000\}$ and $\gamma \in \{0.001,\ldots,20\}$, in increments of 0.0001 for the values of each of these parameters. Then, for each pair (C,γ) in this space, the RMSE was calculated by 5-fold cross-validation. Finally, the pair (C,γ) that yielded the smallest value of the RMSE was chosen, and that pair was used to train the entire training set and generate the final classifier.

With the reduced input dimensions and the optimal values of the parameters, the SVM model for energy consumption of the 175 government-owned buildings in our study was validated. The values of the parameters used in that validation are presented in Table 3. The value obtained for the RMSE using the SVM model with those parameters was 14.0161.

Table 3. Results of parameter optimization using the grid-search method

	C	γ
SVM	10,947.4705	0.0010

3.4. ANN implementation

An artificial neural network (ANN) is a nonlinear machine learning model that consists of a number of interconnected processing elements organized into layers similar to neurons in the human brain (Eom *et al.* 2008). An ANN can be applied to various applications categorized as prediction, classification, and pattern recognition. ANNs have been used for a wide variety of prediction tasks in many different fields of business, industry, and science. The past few years have seen increasing interest in ANNs in different fields of civil engineering (Fazel Zarandi *et al.* 2008; Lee *et al.* 2009). Researchers have also explored the use of ANNs to construct predictive models that are more than just standard regression models.

One advantage of an ANN is that it can handle data sets consisting of an unrestricted number of input and output variables, with no prior assumptions or knowledge about the relationships among the input and output variables (Kalaitzakis *et al.* 2002). However, the primary drawback of an ANN is that considerable time is needed to determine the optimal number of layers and hidden neurons, which requires repetitive trial-and-error tuning processes (Guven *et al.* 2009). Another inherent drawback of ANNs is the need to use a large set of training data to obtain an accurate model (Unbrello *et al.* 2007).

In this study, we used a neural network architecture consisting of a multi-layer perceptron (MLP) with back-propagation to train the ANN model (Rumelhart *et al.* 1986; Bishop 1995). This is arguably the standard neural network model employed to date (Eom *et al.* 2008; Yilmaz, Kaynar 2011). This supervised learning algorithm has certain advantages in regression-type prediction problems. Hornik *et al.* (1990) demonstrated empirically that given the right size and structure, an MLP is capable of learning arbitrarily complex nonlinear functions to an arbitrarily high level of accuracy.

In training an ANN model, two important factors should be considered: the ANN structure and the training iteration number (epoch). Appropriate selection of these two factors prevents the ANN model from becoming overtrained. The MLP used in this study contains one input layer, one hidden layer, and one output layer. The number of nodes in the hidden layer was varied between 2 and 30, and each ANN classifier was constructed using 10,000 epochs as the stopping criterion for training. Several functions can be used as transfer functions, but the most common choice is the sigmoid function (Huang *et al.* 2008). Therefore, a sigmoid transfer function was used as the hidden layer transfer function as well as the output layer transfer function in this study.

3.5. DT implementation

Decision trees can be used for two types of problems: classification (the decision class is a discrete variable – a label or category to which the data belongs) and regression (the decision class is a continuous variable). In this study, we used the M5P algorithm, which was first introduced by Quinlan (1992) and is based on the model tree. The model tree is a special type of decision tree model developed to solve nonlinear regression problems (Quinlan 1992; Wang, Witten 1997). M5P is powerful, because it combines decision trees and linear regression to predict the value of a continuous variable (Quinlan 1992). M5P divides the sample space into many rectangular areas with edges that are approximately parallel. Then it determines a regression model corresponding to each of these rectangular areas. In this respect it has an advantage over multivariate linear regression, since it can approximate nonlinear problems. Another advantage is that the M5P algorithm allows the input to be a mixture of discrete and continuous variables (Quinlan 1992).

The principle behind M5P is fairly simple: it partitions the data into smaller subsets in a decision tree format using training data and their outcomes. Then it fits a linear regression model at each terminal node of the tree by using the data at that node to predict the outcome, instead of directly attaching values to the nodes. Thus, to predict the target value for a given data set, we follow a branch through the tree, beginning at the root node and continuing until a terminal node is reached, and then we apply the corresponding regression model.

The tree is constructed through a binary recursive partitioning method (decision tree induction algorithm). This is an iterative process of splitting the data into partitions by minimizing the variation in the values along each branch. In this process, the standard deviation is used to choose the best split at each node. The goal is to maximize the reduction in the standard deviation by testing the possible splits over the training data that reach a particular node, and then splitting it up further on each branch until only a few instances remain (Hu *et al.* 2007). Then the tree is pruned of unwanted nodes, and a smoothing procedure is applied to avoid sharp discontinuities between adjacent linear models at the leaves of the pruned tree. However, the tree induction of model trees is sequential in nature and locally optimal at each node split. Thus, convergence for a global optimal solution is not always feasible. In addition, minor modifications in the training data could lead to large changes in the final model because of the intrinsic instability of the M5P algorithm (Fan, Gray 2005).

3.6. MLR implementation

Multiple linear regression is a technique used to model a linear relationship among two or more independent variables and a dependent variable. The computational problem addressed by multiple linear regression consists of fitting a plane to an *n*-dimensional space, where *n* is the number of independent variables, as follows:

$$Y = b_0 + \sum_{i=1}^{n} b_i X_i + \varepsilon_i, \tag{10}$$

where: Y is the dependent variable, the X_i are the independent variables, and the constant term b_0 and the regression coefficients b_i are computed by the ordinary least-squares method so that the average error ε is zero (Grivas, Chaloulakou 2006). The regression coefficients represent the amount of change in the dependent variable Y as a result of a change of one unit in the respective independent variables.

The MLR method is based on a few assumptions (Rajaee 2011), namely, that the regression estimators are optimal in the sense that they are unbiased, efficient, and consistent. "Unbiased" means that the expected value of the estimator is equal to the actual value of the parameter. "Efficient" means that the estimator has a smaller variance than any other estimator, and "consistent" means that the bias and variance of the estimator technique go to zero as the sample size approaches infinity.

The major drawback of MLR is its inability to cope with a highly nonlinear problem. In addition, in the regression equation, collinearity between the independent variables can lead to incorrect identification of the most important predictors (Sousa *et al.* 2007). Therefore, the modeling performance of MRL is reportedly poor. However, because of its simplicity, this study investigated the applicability of MLR for comparison purposes.

4. Results and discussion

The performance of the proposed method – the SVM method – was compared to that of three other data mining techniques: ANN, DT, and MLR. Other researchers have proposed using these other three data mining techniques to predict energy consumption in buildings (see, for example, Sharp 1996; Chung *et al.* 2006; Yalcintas 2006; Yalcintas, Ozturk 2007; Lee 2008; Chung, Hui 2009; Yu *et al.* 2010).

In this study, the performance of the predictive models was measured by how closely the predicted values matched the test data and the actual values, as indicated by the prediction errors (that is, the deviations of the predicted values from the actual values). The prediction performance was evaluated by means of a 5-fold cross-validation based on three measures of error: the mean absolute error (MAE), the root mean squared error (RMSE), and the mean absolute percentage error (MAPE):

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - f(x_i)\right|; \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2}; \tag{12}$$

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - f(x_i)}{y_i}\right| \times 100. \tag{13}$$

The three measures listed above are the ones most commonly used (Fan *et al.* 2009; Wen *et al.* 2009). Since the MAE and the RMSE are based on absolute errors, there is no absolute criterion for a "good" value for either of them. All that can be inferred from them is that the smaller the value of MAE or RMSE, the closer the predicted values to the actual values. The MAPE, however, is scale independent, since it is based on relative errors; hence, it is more meaningful (Makridakis 1993).

A comparison of the performance of the four models, as measured by the three aforementioned indices – MSE, RMSE, and MAPE – is presented in Table 4. In order of decreasing performance, the ranking of the four methods with respect to MAE and MAPE is SVM, MLR, ANN, and DT. The corresponding ranking with respect to RMSE is SVM, ANN, DT, and MLR. The proposed SVM model achieves 34.8804% in terms of MAPE, which indicates the highest performance among the four methods. It lies between 20% and 50%, indicating that the predictions made by that model are reasonable (Lewis 1982). Although the proposed model may not be highly accurate, it contributes to project stakeholders' understanding of the effects of design decisions upon the energy performance of the building to be built in the early phases of development of energy-efficient buildings. In other words, the proposed model may be less accurate but is still considered to be highly informative in assisting government agencies with understanding the degree of sustainability of their buildings during the design phase.

Graphical comparisons of the actual target values to the values predicted by the proposed SVM model and the other three models are presented in Figure 2. The values predicted by the SVM model are closer to the actual target values, from lowest to highest, than are those predicted by the other three models. Thus one could conclude that the predicted values show relatively good agreement with the measured values and that the proposed SVM model is feasible and reliable.

From our analysis of the empirical results, it appears that the proposed SVM model performed better than any of the other three models, in that it yielded the smallest MAE, RMSE, and MAPE. At this point, we are interested in doing a more in-depth examination of whether the proposed SVM model is superior to the ANN, DT, and MLR models with regard to prediction of energy consumption of government-owned buildings. We are also interested in determining whether the rankings of the other three models with respect to the three aforementioned measures (MAE, RMSE, and MAPE) have any particular significance in this regard.

Table 4. Comparison to other methods

	MAE	RMSE	MAPE (%)
SVM	12.3333	16.8526	34.8804
ANN	14.5828	18.3704	43.5864
DT	14.7828	18.6746	44.5736
MLR	14.4448	19.4897	41.8701

To test whether the proposed SVR model is superior to the other three models, we applied the two-tailed Wilcoxon matched-pairs signed rank test with respect to RMSE and MAPE. This test was chosen because it is a non-parametric method and it imposes no restriction on the underlying distributions in the data. In addition, it is not performed on the magnitudes of the values of the variables but on their signs and ranks; thus, it is not influenced by outlier data points. It is one of the most widely adopted tests used in evaluating the predictive capabilities of different models to see whether there is a statistically significant difference between them (Pollock *et al.* 2005; Lu *et al.* 2009). Details of the Wilcoxon signed-rank test can be found in Diebold and Mariano (1995) and Pollock *et al.* (2005).

The null hypothesis of the two-tailed Wilcoxon signed-rank test is that the difference between the values of the RMSE (or the MAPE) for the two models being compared is zero. If the performance of one model considerably surpasses the other, those differences will be significantly different from zero. The null hypothesis

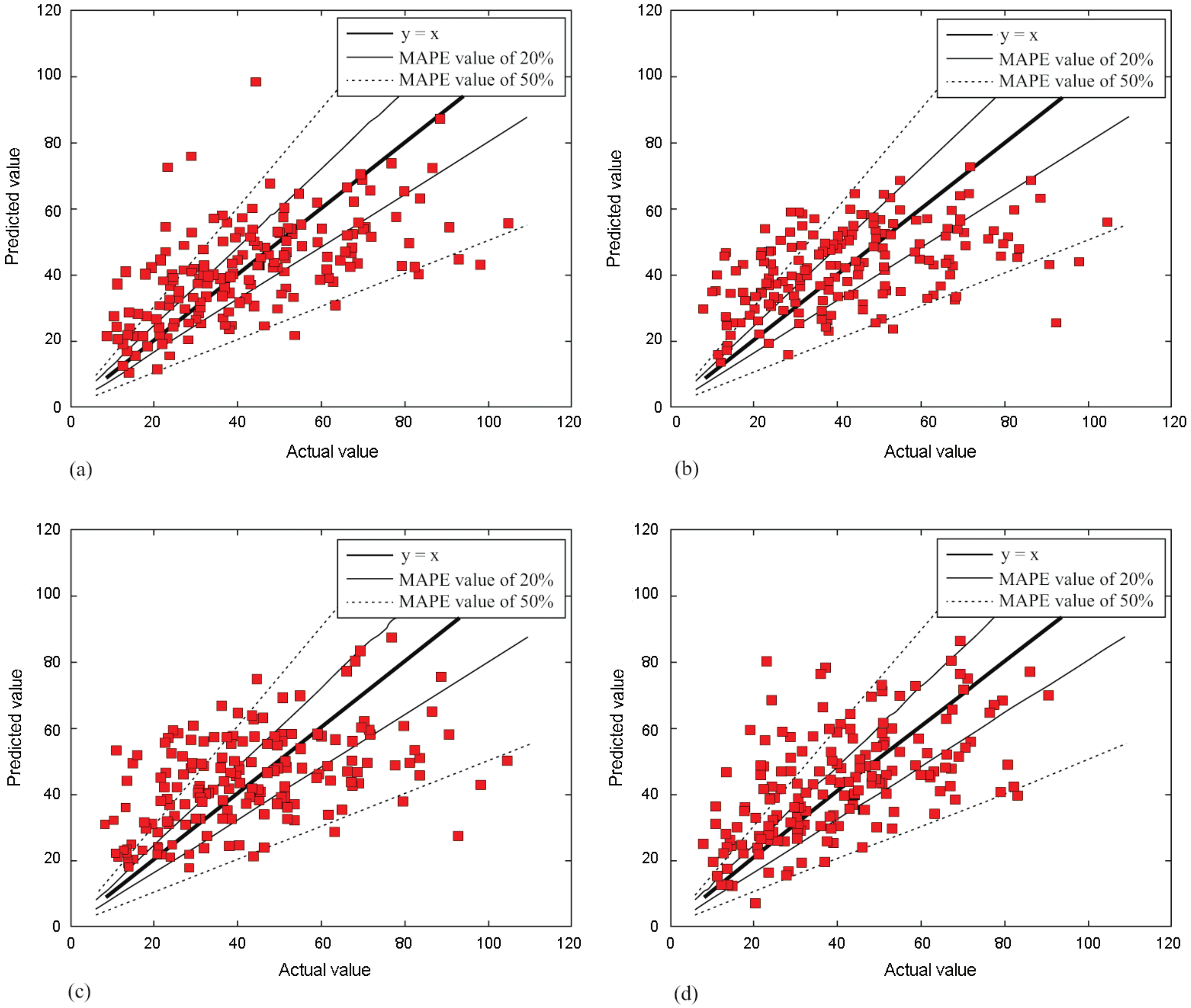

Fig. 2. Comparison of actual values versus predicted values for (a) the proposed SVM model, (b) the ANN model, (c) the DT model, and (d) the MLR model

Table 5. Results of the Wilcoxon signed- rank test with respect to the RMSE

	ANN	DT	MLR
SVM	–3.955[a] (0.000)*	–3.731[a] (0.000)*	–2.873[a] (0.004)*
ANN		–0.162[a] (0.872)	–1.820[b] (0.069)
DT			–1.222[b] (0.222)

[a] based on positive ranks;
[b] based on negative ranks.

Table 6. Results of the Wilcoxon signed-rank test with respect to the MAPE

	ANN	DT	MLR
SVM	–4.158[a] (0.000)*	–3.862[a] (0.000)*	–3.309[a] (0.001)*
ANN		–0.018[b] (0.986)	–1.734[b] (0.083)
DT			–0.919[b] (0.358)

[a] based on positive ranks;
[b] based on negative ranks.

is rejected (meaning that the values of the RMSE and MAPE for the two models are significantly different) when the *p*-value is below a pre-established level of significance. In this study, the minimum significance level for rejecting a null hypothesis was set at 5%.

Tables 5 and 6 present the values of the Z statistic from the two-tailed Wilcoxon signed-rank test with respect to the RMSE and the MAPE, respectively, in the four models. The numbers given in parentheses in those tables are the corresponding *p*-values. The asterisks indicate the *p*-values that are less than 5%. In those two tables, it is shown that the RMSE and the MAPE for the proposed SVM model are significantly different from those for the other three models. They also show that there are no significant differences in the prediction performance among the other three models (i.e. ANN, DT, and MLR). Since the proposed SVM method yielded the smallest RMSE and MAPE values in this study, as well as the best scores on the two-tailed Wilcoxon signed-rank test for those two measures, we can conclude that the prediction performance of the proposed SVM model is significantly better than that of the other three models in regard to energy consumption of government-owned buildings.

Conclusions

This study identified the variables associated with energy consumption in government-owned buildings and proposed a model for predicting the energy consumption of government-owned buildings based on the identified variables. The RReliefF variable selection algorithm identified seven highly relevant variables from a set of 26 candidate variables related to the general building information and energy end uses; building activities and special measures of size; heating and cooling equipment and conservation features; water heating, refrigeration, office equipment, and special uses of space; and lighting percents, equipment, and conservation features. Then the SVM method was used to construct a predictive model of energy consumption based on the seven selected variables. The use of RReliefF for variable selection allowed us to eliminate variables that are irrelevant or redundant, as well as to reduce the dimensionality of the input variables fed to the SVM model. With fewer – and more relevant – input variables, the SVM model was better able to describe the nonlinear relationship between the input variables and the dependent variable, namely, electricity consumption. Consequently, the proposed method was able to more accurately predict the energy consumption of government-owned buildings.

In summary, several interesting findings have been made in this study. First, prior to this study there was relatively little understanding of the variables that significantly contribute to the energy consumption of government-owned buildings. This study identified seven highly relevant variables for predicting the energy consumption of government-owned buildings. The results imply that the accuracy of the proposed model in predicting the energy consumption of a government-owned building is highly dependent on decisions made relative to those variables during the design phase. In other words, the energy consumption of government-owned buildings can be predicted with reasonable accuracy in the design phase based upon the values of these seven variables. Second, the prediction performance obtained using the proposed SVM model was compared with that obtained using three other data-mining techniques that were proposed in previous studies of energy consumption of buildings: ANN, DT, and MLR. The results of the comparison confirmed that the SVM model is the best predictor of the energy consumption of government-owned buildings, as it yields comparatively smaller errors than the others. This reasonably accurate prediction method also has great potential for solving other prediction problems in the construction industry.

The proposed model of energy consumption of government-owned buildings can be utilized to credibly assess the future energy consumption of government-owned buildings before they are built, based on design decisions. Hence, it can be utilized to assist government agencies in understanding the degree of sustainability of their buildings during the design phase. Ultimately, an accurate model of future energy consumption of government-owned buildings will serve as a guide for the early development of strategies to control the consumption of energy, thereby contributing to the sustainability of development in the construction industry.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0023229).

References

Bishop, C. M. 1995. *Neural networks for pattern recognition*. Oxford: Clarendon Press. 482 p.

Chen, M.-Y. 2011. Predicting corporate financial distress based on integration of decision tree classification and logistic regression, *Expert Systems with Applications* 38(9): 11261–11272. http://dx.doi.org/10.1016/j.eswa.2011.02.173

Chung, W.; Hui, Y. V.; Lam, Y. M. 2006. Benchmarking the energy efficiency of commercial buildings, *Applied Energy* 83(1): 1–14. http://dx.doi.org/10.1016/j.apenergy.2004.11.003

Chung, W.; Hui, Y. V. 2009. A study of energy efficiency of private office buildings in Hong Kong, *Energy and Buildings* 41(6): 696–701. http://dx.doi.org/10.1016/j.enbuild.2009.02.001

Diebold, F. X.; Mariano, R. S. 1995. Comparing predictive accuracy, *Journal of Business and Economic Statistics* 13(3): 253–263.

Eom, J.-H.; Kim, S.-C.; Zhang, B.-T. 2008. AptaCDSS-E: a classifier ensemble-based clinical decision support system for cardiovascular disease level prediction, *Expert Systems with Applications* 34(4): 2465–2479. http://dx.doi.org/10.1016/j.eswa.2007.04.015

European Commission (EC). 2012. *Energy efficiency in buildings* [online], [cited 20 September 2012]. Available from Internet: http://ec.europa.eu/energy/efficiency/buildings/buildings_en.htm

Fan, G.; Gray, J. B. 2005. Regression tree analysis using target, *Journal of Computational and Graphical Statistics* 14(1): 206–218. http://dx.doi.org/10.1198/106186005X37210

Fan, S.; Liao, J. R.; Yokoyama, R.; Chen, L.; Lee, W.-J. 2009. Forecasting the wind generation using a two-stage network based on meteorological information, *IEEE Transactions on Energy Conversion* 24(2): 474–482. http://dx.doi.org/10.1109/TEC.2008.2001457

Fazel Zarandi, M. H.; Türksen, I. B.; Sobhani, J.; Ramezanianpour, A. A. 2008. Fuzzy polynomial neural networks for approximation of the compressive strength of concrete, *Applied Software Computing* 8(1): 488–498. http://dx.doi.org/10.1016/j.asoc.2007.02.010

Grivas, G.; Chaloulakou, A. 2006. Artificial neural network models for prediction of PM10 hourly concentrations in the Greater Area of Athens, Greece, *Atmospheric Environment* 40(7): 1216–1229. http://dx.doi.org/10.1016/j.atmosenv.2005.10.036

Guven, A.; Azamathulla, H. Md.; Zakaria, N. A. 2009. Linear genetic programming for prediction of circular pile scour, *Ocean Engineering* 36(12–13): 985–991. http://dx.doi.org/10.1016/j.oceaneng.2009.05.010

Han, Y.; Yu, L. 2012. A variance reduction framework for stable feature selection, *Statistical Analysis and Data Mining* 5(5): 428–445. http://dx.doi.org/10.1002/sam.11152

Hornik, K.; Stinchcombe, M.; White, H. 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, *Neural Networks* 3(5): 551–560. http://dx.doi.org/10.1016/0893-6080(90)90005-6

Huang, S.-C. 2011. Forecasting stock indices with wavelet domain kernel partial least square regressions, *Applied Soft Computing* 11(8): 5433–5443. http://dx.doi.org/10.1016/j.asoc.2011.05.015

Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. 2003. *A practical guide to support vector classification*. Technical Report. National Taiwan University, Taipei, Taiwan.

Hu, P. J.-H.; Cheng, T.-H.; Wei, C.-P.; Yu, C.-H.; Chan, A. L. F.; Wang, H.-Y. 2007. Managing clinical use of high-alert drugs: a supervised learning approach to pharmacokinetic data analysis, *IEEE Transactions on Systems, Man, and Cybernetics* 37(4): 481–492. http://dx.doi.org/10.1109/TSMCA.2007.897700

Huang, C.-J.; Yang, D.-X.; Chuang, Y.-T. 2008. Application of wrapper approach and composite classifier to the stock trend prediction, *Expert Systems with Applications* 34(4): 2870–2878. http://dx.doi.org/10.1016/j.eswa.2007.05.035

International Energy Agency (IEA). 2010. *Energy performance certification of buildings: a policy tool to improve energy efficiency* [online]. The IEA Policy Pathway Series, International Energy Agency, Paris, France [cited 20 September 2012]. Available from Internet: http://www.iea.org/papers/pathways/buildings_certification.pdf

Kalaitzakis, K.; Stavrakakis, G. S.; Anagnostakis, E. M. 2002. Short-term load forecasting based on artificial neural networks parallel implementation, *Electric Power Systems Research* 63(3): 185–196. http://dx.doi.org/10.1016/S0378-7796(02)00123-2

Kandaswamy, K. K.; Pugalenthi, G.; Hazrati, M. K.; Kalies, K.-U.; Martinetz, T. 2011. BLProt: prediction of bioluminescent proteins based on support vector machine and relieff feature selection, *BMC Bioinformatics* 12(17): 1–7.

Kira, K.; Rendell, L. A. 1992. A practical approach to feature selection, in *Proc. of the 9th International Workshop on Machine Learning*, 12–16 July 1992, San Francisco, CA, 249–256.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, in *Proc. of the International Joint Conference on Artificial Intelligence*, 20–25 August 1995, Québec, Canada, 1137–1143.

Kononenko, I. 1994. Estimating attributes: analysis and extensions of Relief, *Machine Learning* 784: 171–182.

Kononenko, I.; Robnik-Šikonja, M.; Pompe, U. 1996. ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems, in *Proc. of the International Conference on Artificial Intelligence: Methodology, Systems, Applications*, 18–20 September 1996, Sozopol, Bulgaria, 31–40.

Kumar, K.; Bhattacharya, S. 2006. Artificial neural network vs linear discriminant analysis in credit ratings forecast: a comparative study of prediction performances, *Review of Accounting and Finance* 5(3): 216–227. http://dx.doi.org/10.1108/14757700610686426

Lee, J. J.; Kim, D.; Chang, S. K.; Nocete, C. F. M. 2009. An improved application technique of the adaptive probabilistic neural network for predicting concrete strength, *Computational Materials Science* 44(3): 988–998. http://dx.doi.org/10.1016/j.commatsci.2008.07.012

Lee, W.-S. 2008. Benchmarking the energy efficiency of government buildings with data envelopment analysis, *Energy and Buildings* 40(5): 891–895. http://dx.doi.org/10.1016/j.enbuild.2007.07.001

Lewis, C. D. 1982. *International and business forecasting methods*. London, UK: Butterworths. 40 p.

Li, H.; Sun, J. 2011. Predicting business failure using support vector machines with straightforward wrapper: a resampling study, *Expert Systems with Applications* 38(10): 12747–12756. http://dx.doi.org/10.1016/j.eswa.2011.04.064

Lu, C.-J.; Lee, T.-S.; Chiu, C.-C. 2009. Financial time series forecasting using independent component analysis and support vector regression, *Decision Support Systems* 47(2): 115–125. http://dx.doi.org/10.1016/j.dss.2009.02.001

Makridakis, S. 1993. Accuracy measures: theoretical and practical concerns, *International Journal of Forecasting* 9(4): 527–529. http://dx.doi.org/10.1016/0169-2070(93)90079-3

Molina, L. C.; Belanche, L.; Nebot, A. 2002. Feature selection algorithms: a survey and experimental evaluation, in *Proc. of the 2002 IEEE International Conference on Data Mining*, 9–12 December 2002, Maebashi City, Japan, 306–313.

Moore, D. S.; McCabe, G. P. 1999. *Introduction to the practice of statistics*. 3rd ed. New York, NY: W.H. Freeman. 825 p.

Pernek, I.; Stigli, G.; Kokol, P. 2012. How hard am I training? Using smart phones to estimate sport activity intensity, in *Proc. of the 32nd International Conference on Distributed Computing Systems Workshops*, 18–21 June 2012, Macau, China, 65–68.

Pham, D. T.; Castellani, M.; Fahmy, A. A. 2009. Evolutionary feature selection for artificial neural network pattern classifiers, in *Proc. of the 7th IEEE International Conference on Industrial Informatics*, 24–26 June 2009, Wales, UK, 658–663.

Pollock, A. C.; Macaulay, A.; Thomson, M. E.; Onkal, D. 2005. Performance evaluation of judgemental directional exchange rate predictions, *International Journal of Forecasting* 21(3): 473–489. http://dx.doi.org/10.1016/j.ijforecast.2004.12.006

Quinlan, R. J. 1992. Learning with continuous classes, in *Proc. of the 5th Australian Joint Conference on Artificial Intelligence*, 16–18 November 1992, Hobart, Tasmania, 343–348.

Rajaee, T. 2011. Wavelet and ANN combination model for prediction of daily suspended sediment load in rivers, *Science of the Total Environment* 409(15): 2917–2928. http://dx.doi.org/10.1016/j.scitotenv.2010.11.028

Robnik-Šikonja, M.; Kononenko, I. 1997. An adaptation of Relief for attribute estimation in regression, in *Proc. of the 14th International Conference on Machine Learning*, 8–12 July 1997, Nashville, TN, 296–304.

Robnik-Šikonja, M.; Kononenko, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF, *Machine Learning* 53(1–2): 23–69. http://dx.doi.org/10.1023/A:1025667309714

Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. 1986. *Learning internal representation by error propagation. Parallel distributed processing*. Cambridge, MA: MIT Press, 318–362.

Salzberg, S. L. 1997. On comparing classifiers: Pitfalls to avoid and a recommended approach, *Data Mining and Knowledge Discovery* 1(3): 317–328. http://dx.doi.org/10.1023/A:1009752403260

Schölkopf, B.; Burges, C. J. C.; Smola, A. J. 1999. *Advances in Kernel methods: Support vector learning*. Cambridge, MA: MIT Press. 376 p.

Sharp, T. 1996. Energy benchmarking in commercial office buildings, in *Proc. of the ACEEE 1996 Summer Study on Energy Efficiency in Buildings*, 25–31 August 1996, Washington, DC, 321–329.

Sousa, S. I. V.; Martins, F. G.; Alvim-Ferraz, M. C. M.; Pereira, M. C. 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations, *Environmental Modelling & Software* 22(1): 97–103. http://dx.doi.org/10.1016/j.envsoft.2005.12.002

Srisawat, A.; Kijsirikul, B. 2009. Predicting HIV-1 drug resistance: a comparison of three learning algorithms, in *Proc. of the 2nd International Conference on Biomedical Engineering and Informatics*, 17–19 October 2009, Tianjin, China, 1–5.

Suykens, J. A. K.; Vandewalle, J. 1998. *Nonlinear modeling: advanced black-box techniques*. Boston, MA: Kluwer Academic. 256 p. http://dx.doi.org/10.1007/978-1-4615-5703-6

U.S. Department of Energy (DOE). 2012. *Buildings energy data book* [online], [cited 20 September 2012]. Available from Internet: http://buildingsdatabook.eren.doe.gov

U.S. National Archives and Records Administration (NARA). 2007. *Executive Order 13423* [online], [cited 20 September 2012]. Available from Internet: http://edocket.access.gpo.gov/2007/pdf/07-374.pdf

Unbrello, D.; Ambrogio, G.; Filice, L.; Shivpuri, R. 2007. An ANN approach for predicting subsurface residual stresses and the desired cutting conditions during hard turning, *Journal of Materials Processing Technology* 189(1–3): 143–152. http://dx.doi.org/10.1016/j.jmatprotec.2007.01.016

Vapnik, V.; Lerner, A. 1963. Pattern recognition using generalized portrait method, *Automation and Remote Control* 24: 774–780.

Vapnik, V. N. 1995. *The nature of statistical learning theory*. New York, NY: Springer-Verlag. 188 p. http://dx.doi.org/10.1007/978-1-4757-2440-0

Wang, J.; Wu, X.; Zhang, C. 2005. Support vector machines based on K-means clustering for real-time business intelligence systems, *International Journal of Business Intelligence and Data Mining* 1(1): 54–64. http://dx.doi.org/10.1504/IJBIDM.2005.007318

Wang, W.; Xu, Z.; Lu, J. W. 2003. Three improved neural network models for air quality forecasting, *Engineering Computations* 20(2): 192–210. http://dx.doi.org/10.1108/02644400310465317

Wang, Y.; Witten, I. H. 1997. Inducing model trees for continuous classes, in *Proc. of the Poster Papers of the 9th European Conference on Machine Learning*, 23–25 April 1997, Prague, Czech Republic, 128–137.

Wen, Y. F.; Cai, C. Z.; Liu, X. H.; Pei, J. F.; Zhu, X. J.; Xiao, T. T. 2009. Corrosion rate prediction of 3C steel under different seawater environment by using support vector regression, *Corrosion Science* 51(2): 349–355. http://dx.doi.org/10.1016/j.corsci.2008.10.038

Yalcintas, M. 2006. An energy benchmarking model based on artificial neural network method with a case example for tropical climates, *International Journal of Energy Research* 30(14): 1158–1174. http://dx.doi.org/10.1002/er.1212

Yalcintas, M.; Ozturk, U. A. 2007. An energy benchmarking model based on artificial neural network method utilizing US Commercial Buildings Energy Consumption Survey (CBECS) database, *International Journal of Energy Research* 31(4): 412–421. http://dx.doi.org/10.1002/er.1232

Yang, Y.-H.; Lin, Y.-C.; Su, Y.-F.; Chen, H. H. 2008. A regression approach to music emotion recognition, *IEEE Transactions on Audio, Speech, and Language Processing* 16(2): 448–457.

Yilmaz, I.; Kaynar, O. 2011. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils, *Expert Systems with Applications* 38(5): 5958–5966. http://dx.doi.org/10.1016/j.eswa.2010.11.027

Yu, Z.; Haghighat, F.; Fung, B. C. M.; Yoshino, H. 2010. A decision tree method for building energy demand modeling, *Energy and Buildings* 42(10): 1637–1646. http://dx.doi.org/10.1016/j.enbuild.2010.04.006

Hyojoo SON. She is a Researcher at Chung-Ang University. She holds MS in Architectural Engineering from Chung-Ang University. Her main research interests include data mining in construction management, sensor-based remote intelligent monitoring of construction jobsites, computer vision, and information modeling in civil and infrastructure engineering.

Changmin KIM. He is a Master's student at Chung-Ang University. His Master's research focuses on structural equation modeling and its application within the field of civil engineering.

Changwan KIM. He is an Associate Professor of Architectural Engineering at Chung-Ang University, Korea. His works appeared in publications such as *Expert Systems with Applications*, *Transportation Research Part E*, *Automation in Construction*, *Computing in Civil Engineering*, and *Sustainable Development*. His research interests include knowledge discovery and data-mining for civil and infrastructure engineering, sensor and sensing for intelligent project management, construction automation, intelligent building and construction automation.

Youngcheol KANG. He is an Assistant Professor, at the Department of Global Construction Management, the University of Seoul. His papers appeared in journals such as *Automation in Construction*, *Journal of Computing in Civil Engineering*, and *Journal of Construction Engineering and Management*. His research interests are primarily involved in project management related to knowledge discovery in databases, quantitative methods, decision, risk and reliability, cost management, and sustainable development.