Available online at www.sciencedirect.com
SCIENCE DIRECT®
Energy and Buildings 37 (2005) 545–553
ENERGY and BUILDINGS
www.elsevier.com/locate/enbuild

Applying support vector machines to predict building energy consumption in tropical region

Bing Dong[a,*], Cheng Cao[b], Siew Eang Lee[a]

[a]*Department of Building, School of Design and Environment, National University of Singapore, 4 Architecture Drive, 117566 Singapore, Singapore*
[b]*Department of Mechanical Engineering, School of Engineering, National University of Singapore, 9 Engineering Drive, 117596 Singapore, Singapore*

Received 17 August 2004; received in revised form 17 September 2004; accepted 25 September 2004

Abstract

The methodology to predict building energy consumption is increasingly important for building energy baseline model development and measurement and verification protocol (MVP). This paper presents support vector machines (SVM), a new neural network algorithm, to forecast building energy consumption in the tropical region. The objective of this paper is to examine the feasibility and applicability of SVM in building load forecasting area. Four commercial buildings in Singapore are selected randomly as case studies. Weather data including monthly mean outdoor dry-bulb temperature (T_0), relative humidity (RH) and global solar radiation (GSR) are taken as three input features. Mean monthly landlord utility bills are collected for developing and testing models. In addition, the performance of SVM with respect to two parameters, C and ε, was explored using stepwise searching method based on radial-basis function (RBF) kernel. Finally, all prediction results are found to have coefficients of variance (CV) less than 3% and percentage error (%error) within 4%.

Keywords: Building energy consumption prediction; Support vector machine; Weather data; Tropical region

1. Introduction

Previous researches have shown that baseline building energy consumption is one of the key elements for energy performance contracting and measurement and verification protocol (MVP). A baseline model of this kind can directly decide the energy savings that could occur after a building undergoes retrofitting works. Many efforts have been taken on the development and improvement of the baseline models as accurate as possible. Fels [1] utilized variable-base degree-day method to estimate residential retrofitting energy use. Kissock [2] developed a regression methodology to measure retrofitting energy use in commercial buildings. Krarti et al. [3] utilized neural networks to estimate energy and demand savings from retrofits of commercial buildings. Dhar et al. [4] generalized the Fourier series approach to model hourly energy use in commercial buildings. However, in most practical cases, utility bill data are used because they are widely available and inexpensive to obtain and process. In the tropical region, Dong et al. [5,6] utilized two years' utility bills to establish a baseline model and the results showed small percentage error (%error) with high prediction intervals.

It is well know that for the building energy consumption, there are certain constant loads, such as lighting, fan and plug loads, which do not change with weather variables, especially the landlord energy consumption. Landlord energy consumption refers to the energy utilized inside the part of the building, typically comprising: (a) air-conditioner central plant system which supply air-conditioning inside the building; (b) vertical transportation service such as escalator and lift; (c) ventilation system such as exhaust fan and ventilator; (d) artificial lighting system in the common area, i.e. corridor or public common service areas such as toilets and lifts. Obviously, the usages of these

* Corresponding author. Tel.: +65 68745891; fax: +65 67733837.
E-mail address: g0203869@nus.edu.sg (B. Dong).

0378-7788/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.enbuild.2004.09.009

systems present certain non-linear performances between building energy use and weather data. Furthermore, the building owner often received the landlord bills only. It seems more meaningful to baseline landlord energy consumption rather than the whole building energy use for the benefits of both building owners and ESCOs in the energy efficiency measures (ECMs) retrofitting contracts. Hence, these practical reasons motivated us to make an exploration of non-linear performance for estimating landlord energy consumption. Dong et al. [5,6] have tried neural networks to predict the same kind of building energy consumption. However, the prediction results showed low %error with high coefficient of variance (CV) around 20%.

This paper presents a novel neural network technique called support vector machines (SVM) to estimate landlord energy consumption. SVM, developed by Vapnik and his coworkers in 1995, has been widely applied in classification, forecasting and regression ([7–12], Muller et al., 1999). Their practical success can be attributed to solid theoretical foundations based on VC-theory [13]. The objectives of this study are firstly, to examine the feasibility of applying SVM in predicting building energy consumption, and secondly, to investigate the effect of different SVM parameters on the prediction accuracy. Four commercial buildings are selected randomly from the office buildings situated in the central business area for detailed case studies. The whole building landlord energy consumption is chosen to building and test models. The weather parameters, namely, outdoor dry-bulb temperature (T_0), relative humidity (RH) and global solar radiation (GSR) are taken as three features.

The structure of the paper is organized as follows: Section 2 provides a brief introduction to SVM regression; Section 3 applies SVM to four real buildings; the technique used for data processing and statistical performance metrics are described in Section 3 and lastly, the conclusions are presented in Section 4.

2. Support vector machines for regression estimation

Support vector machines (SVM) are based on the structural risk minimization (SRM) inductive principle, which seeks to minimize an upper bound of the generalization error consisting of the sum of the training error and a confidence level. This is the difference from commonly used empirical risk minimization (ERM) principle, which only minimizes the training error. Based on such induction principle, SVM usually achieves higher generalization performance than the traditional neural networks that implement the ERM principle in solving many machine learning problems. Another key characteristic of SVM is that training SVM is equivalent to solving a linearly constrained quadratic programming problem so that the solution of SVM is always unique and globally optimal, unlike other network's training which requires non-linear optimization with the danger of getting stuck into local minima. In SVM, the solution to the problem is only dependent on a subset of training data points which are referred to as support vectors. Using only support vectors, the same solution can be obtained as using all the training data points. One disadvantage of SVM is that the training time scales somewhere between quadratic and cubic with respect to the number of training samples. So a large amount of computation time will be involved when SVM is applied for solving large-size problems [12]. However, in this study, a small data pool is considered.

2.1. Theory of SVM for regression

Given a set of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)$ ($x_i \in X \subseteq R^n, y_i \in Y \subseteq R, l$ is the number of training samples) are randomly and independently generated from an unknown function. SVM approximates the function using the following form [12]:

$$f(x) = \omega \cdot \phi(x) + b \tag{1}$$

where $\phi(x)$ represents the high-dimensional feature spaces which are nonlinearly mapped from the input space x. The coefficients ω and b are estimated by minimizing the regularized risk function (2)

$$\frac{1}{2}||\omega||^2 + C\frac{1}{l}\sum_{i=1}^{l} L_\varepsilon(y_i, f(x_i)) \tag{2}$$

The first term $||\omega||^2$ is called the regularized term. Minimizing $||\omega||^2$ will make a function as flat as possible, thus playing role of controlling the function capacity. The second term $\frac{1}{l}\sum_{i-1}^{l} L_\varepsilon(y_i, f(x_i))$ is the empirical error measured by the ε-insensitive loss function, which is defined below [14];

$$L_\varepsilon(y_i, f(x_i)) = \begin{cases} |y_i - f(x_i)| - \varepsilon, |y_i - f(x_i)| \geq \varepsilon \\ 0 \end{cases} \tag{3}$$

This defines a ε tube (Fig. 1) so that if the predicted value is within the tube the loss is zero, while if the predicted point is

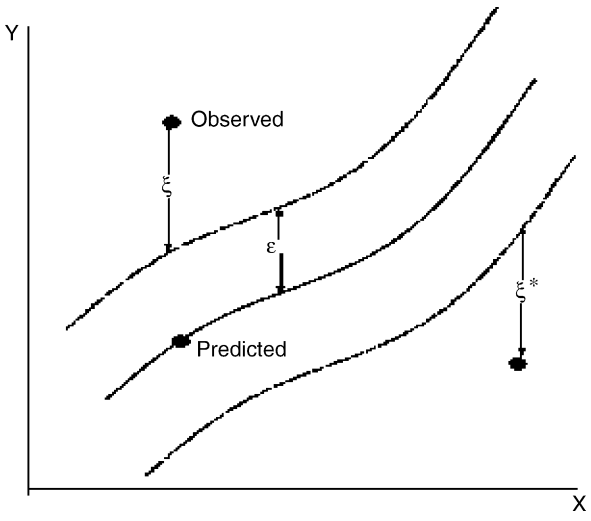

Fig. 1. The parameters for the support vector regression [14].

outside the tube, the loss is magnitude of the difference between the predicted value and the radius ε of the tube. C is called the regularization constant. They are both determined by users.

To get the estimation of ω and b, Eq. (2) is transformed to the primal objective function (4) by introducing the positive slack variables ξ_i^* ((*) denotes variables with and without *).

$$\text{Minimize } \frac{1}{2}||\omega||^2 + C\frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$

Subject to

$$\begin{aligned} &y_i - \omega\cdot\phi(x_i) - b \le \varepsilon + \xi_i \\ &\omega\cdot\phi(x_i) + b - y_i \le \varepsilon + \xi_i^*, \quad i = 1,\ldots,l \\ &\xi_i^* \ge 0 \end{aligned} \tag{4}$$

In most cases the optimization problem (4) can be solved more easily in its dual formulation (5).

$$\begin{aligned} L = {} & \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi^*) \\ & -\sum_{i=1}^{l}a_i(\varepsilon + \xi_i - y_i + \omega\cdot\phi(x_i) + b) \\ & -\sum_{i=1}^{l}a_i^*(\varepsilon + \xi_i^* - y_i - \omega\cdot\phi(x_i) + b) \end{aligned} \tag{5}$$

Here L is the Lagrangian and $\eta_i, \eta_i^*, a_i, a_i^*$ are Lagrange multipliers. Hence, the dual variables in (5) have to satisfy positive constraints,

$$\eta_i^*, a_i^* \ge 0 \tag{6}$$

Again, we refer to a_i^* ((*) denotes variables with and without *).

It follows from the saddle point condition that the partial derivatives of L with respect to the primal variables.

$$\partial_b L = \sum_{i=1}^{l}(a_i + a_i^*) = 0 \tag{7}$$

$$\partial_\omega L = \omega - \sum_{i=1}^{l}(a_i + a_i^*)\phi(x_i) = 0 \tag{8}$$

$$\partial_{\xi_i^*} L = C - a_i^* - \eta^* = 0 \tag{9}$$

Substituting (7), (8), and (9) in (5) yields the dual optimization problem.

$$\begin{aligned} &\text{Maximize } W(a_i, a_i^*) \\ &= \sum_{i=1}^{l}y_i(a_i - a_i^*) - \varepsilon\sum_{i=1}^{l}(a_i + a_i^*) \\ &\quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i - a_i^*)(a_j - a_j^*)(\phi(x_i)\cdot\phi(x_j)) \end{aligned} \tag{10}$$

$$\text{Subject to} \sum_{i=1}^{l}(a_i - a_i^*) = 0 \quad \text{and} \quad a_i, a_i^* \in [0, C];$$

Through condition (8) we can get

$$\omega - \sum_{i=1}^{l}(a_i - a_i^*)x_i = 0$$

thus

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*)\phi(x_i)\cdot\phi(x) + b. \tag{11}$$

By introducing kernel function $K(x_i,x_j)$ the Eq. (11) can be rewritten as follows:

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*)K(x_i, x) + b \tag{12}$$

The value of $K(x_i,x_j)$ is equal to the inner product of two vectors x_i and x_j in the feature space $\phi(x_i)$ and $\phi(x_j)$, that is, $K(x_i,x_j) = \phi(x_i)\cdot\phi(x_j)$. By the use of kernels, all necessary computations can be performed directly in input space, without having to compute the map $\phi(x)$. Some popular kernel functions are the linear kernel $K(x_i,x_j) = x_i\cdot x_j$, polynomial kernel $K(x_i,x_j) = (x_i\cdot x_j + 1)^d$ and the radial-basis function (RBF) kernel $K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$, $\gamma>0$, where d and γ are the kernel parameters. Using different kernel functions, one can construct different learning machines with arbitrary types of decision surfaces.

Based on the Karush–Kuhn–Tucker (KKT) conditions of quadratic programming, only a certain number of coefficients $(a_i - a_i^*)$ in Eq. (5) will assume non-zero values. The data points associated with them have approximation errors equal to or larger than ε and are referred to as support vectors. These are the data points lying on or outside the ε-bound decision function. According to Eq. (5), it is evident that support vectors are the only elements of the data points that are used in determining the decision function as the coefficients $(a_i - a_i^*)$ of other data points are all equal to zero. Generally, the larger the ε, the fewer the number of support vectors and thus the sparser the representation of the solution. However, a larger ε can also depreciate the approximation accuracy placed on the training points. In this sense, ε is a trade-off between the sparseness of the representation and closeness to the data [12].

2.2. Features of SVM for regression estimation

Based on the theoretical statement of SVM, several features of SVM are listed below. Firstly, SVM estimates the regression using a set of linear functions that are defined in a high-dimensional feature space, while the inputs have nonlinear performance. These functions are called kernel functions. Secondly, SVM carries out the regression estimation by risk minimization, based on statistical learning theory, where the risk is measured using Vapnik's ε-insensitive loss function. Finally, SVM implements the SRM principle, which minimizes the risk function consisting of the empirical error and the value of confidence level.

2.3. Necessity of scaling features

Scaling features before applying SVM is very important as well as when neural networks are applied ([15], Part 2 of neural networks FAQ). The main advantage is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. Large attribute values might cause numerical problems because kernel values usually depend on the inner products of feature vectors such as the linear kernel and the polynomial kernel. It is recommended to linearly scale each attribute to the range [−1, +1] or [0, 1]. Likewise, before testing, the same way is applied to scale testing data. For example, it is assumed that one scaled the first attribute of training data from [−10, +10] to [−1, +1]. If the first attribute of testing data is lying in the range [−11, +8], one should scale the testing data to [−1.1, +0.8].

3. Applying SVM in prediction of building energy consumption

3.1. Data collection

Four buildings were selected randomly among all the buildings around the Central Business District in Singapore. They are all peer office buildings for commercial use. The utility bills of these four buildings were collected through surveys, which were carried by the earlier research on building efficiency [16]. In order to retain the individual building anonymity, these four buildings are referred to as Buildings A–D. The first survey began in October 1996 and the survey period lasted for two years, while the second one started in 2000 and ended in 2001. Correspondingly, the whole period of all landlord utility bills is separated into two. As a matter of fact, the electricity consumption data were read from main meters by the building owners as mentioned in [5,6]. Table 1 shows the building size and the average annual energy use of these four buildings. For all the four buildings, the period from October 1996 to October 1998 and year 2000 were selected as training year, while year 2001 was used as the test year.

For the required input features, many previous researches [3,5,6,17,25] have shown that T_0, RH and GSR are three important parameters that affect the building energy consumption. The corresponding weather data are taken from National Environment Agency, Singapore. There are four weather stations in Singapore, namely, Tengah, ChangGi, Seletar and Senbawang. The station in Seletar is selected as it is nearest to four buildings among four stations because the onsite weather data are not available. The monthly data are found by averaging the hourly data of the whole month. Fig. 2 shows the monthly mean weather data during the whole training and testing period. It also shows that there is little difference between month mean temperatures among four years. However, RH and GSR show some variations. For example, the highest GSR is 530.54 cal/cm^2 in February1998, while the lowest point is 317.3 cal/cm^2 in December 2000.

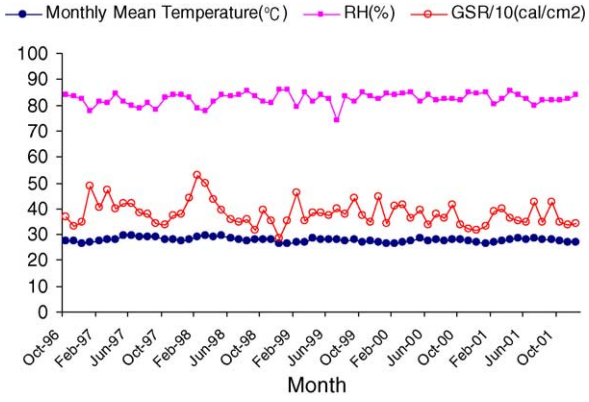

Fig. 2. Weather data profile of five years.

In addition, there is a need to remove the effects of year-to-year changes in conditioned areas and occupancies. Fels and Keating [18] assumed a proportional relationship between annual daily energy use and changes in conditioned area. Hence, normalizing area-changed energy use is merely the annual mean monthly energy use divided by the conditioned area for that particular year. In this study, only landlord energy consumption is considered as it potentially removes the effects of population changes. Hence, we assumed that normalizing energy use by landlord area would be implicitly enough for both changes.

3.2. Utility bill period

As discussed in 2004 [5,6], the utility bill collection period has three possibilities, which is said that utility bill's period could be the same month, previous month and 15 days shift of correspondent weather data period. In order to find out the best prediction model, all possibilities are conducted

Table 1
Size and annual energy consumption of four buildings

Building	Training year	Test year	Total building area (m^2)	Landlord area (m^2)	Total landlord energy consumption (MW/yr)
A	October 1996 to October 1998 and 2000	2001	36,629	14,938	5291
B	October 1996 to October 1998 and 2000	2001	46,400	22,627	6024
C	October 1996 to October 1998 and 2000	2001	60,895	16,910	7681
D	October 1996 to October 1998 and 2000	2001	108,000	41,364	1283

Table 2
Results of utility bill period identification

Building	Bill and weather period
A	Previous
B	15 days shift
C	Same
D	15 days shift

by regression analysis of temperature and landlord energy consumptions before training. The results are shown in Table 2. The utility bill period is the same as weather data period in Building C and is the previous period of that in Building A, while 15 days shift in Building D. After such kind of identification, RH and GSR periods are selected coincide with the temperature period accordingly.

3.3. Model identification

As discussed in Section 2, the selection of kernel function, parameter C, γ and ε directly decides the performance of the final prediction model. A stepwise approaching, which selects the fined parameters directly from the training data, is developed to decide the prediction model.

3.3.1. Performance criteria

The criterion used to select the most appropriate model is to maximize the goodness-of-fit using the simplest model or combination of models [27]. For the non-linear modelling, the prediction performance is evaluated using the following statistical metrics, namely, mean squared error (MSE), mean squared error of scaled value (S-MSE), percentage error (%error) and coefficient of variance based on root mean squared error (CV-RMSE). The CV-RMSE is defined below:

$$\text{CVRMSE} = \frac{\text{RMSE}}{\bar{Y}} \times 100 \tag{13}$$

where

$$\text{RMSE} = [\text{MSE}]^{1/2} = \left[\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})}{n-p}\right]^{1/2} \tag{14}$$

In addition, %error is simply defined below,

$$\%\text{error} = \frac{Y - \hat{Y}}{Y} \times 100 \tag{15}$$

$\hat{Y}$ is the value of Y predicted by the prediction model, n the number of observations and p is the number of model parameters. The smaller the MSE and %error value, the closer are the predicted values to the actual values. CV-RMSE is a non-dimensional measure that is found by dividing RMSE by the mean value of Y. It is usually presented as a percentage. Hence, a CV-RMSE value of 5% would indicate that the mean variation in Y not explained by the regression model is only 5% of the mean value of Y [19]. In addition, during the model identification period, S-MSEs are compared only and enough to decide the best model.

3.3.2. Kernel selection

Most of the previous research selected Gaussian function which is included in RBF as the kernel model for regression.The RBF kernel nonlinearly maps samples into a higher dimensional space, and unlike the linear kernel, can handle the case when the relation between class labels and attributes is non-linear. Furthermore, the linear kernel is a special case of RBF as Keerthi and Lin [20] showed that the linear kernel with a penalty parameter C had the same performance as the RBF kernel with some parameters (C, γ). In addition, the sigmoid kernel behaves like RBF for certain parameters [21]. Another reason can be the number of hyperparameters, which influences the complexity of model selection. The polynomial kernel has more hyperparameters than the RBF kernel. Hence, the RBF kernel has less numerical difficulties in contrast to polynomial kernels whose values may go to infinity or zero. Moreover, it is noted that the sigmoid kernel is not valid (i.e. not the inner product of two vectors) under some parameters [14]. Therefore, RBF kernel is selected in this study. Besides, according to the definition of $-\gamma = 1/k$ by Limsvm-2.6 [22], where k means the number of attributes in the input data, $-\gamma$ is constantly set to 1/3 in the future modelling. Finally, all the training and test data sets are scaled to [0, 1].

3.3.3. Modification of parameter C and ε

There are two parameters except γ while using RBF kernels: C and ε. As discussed before, ε is the key parameter in the ε-insensitive loss function. It is not known beforehand which C and ε are the best choice for the problem; consequently, some kind of model selection (parameter search) should be investigated. The goal is to identify good (C, ε) so that the regressor can accurately predict unknown data such as testing data. Normally, to precisely reflect the performance on regressing unknown data and prevent the over-fitting problem, the cross-validation approach is used. In v-fold cross-validation, the training set is divided into v subsets of equal size. Sequentially, one subset is tested using the regressor trained on the remaining $(v-1)$ subsets. Thus, each instance of the whole training set is predicted once the accuracy of cross-validation is shown as the average MSE. In this study, three-fold validation is conducted. In addition to properly select parameters C and ε, one-time search developed by Tay and his co-workers [12] was firstly performed, and then, a new search method named stepwise search is developed.

3.3.3.1. Selection of parameter C. According to Eq. (4), parameter C determines the trade-off between the model complexity and the degree to which deviations larger than ε are tolerated in optimization formulation (4). In addition, referring to Eq. (10), the regularization parameter C decides the range of values $0 \leq (a_i, a_i^*) \leq C, i = 1, \ldots, l$, assumed

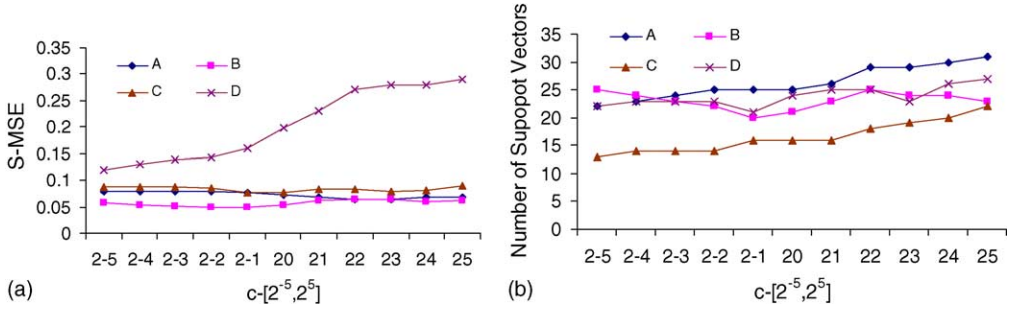

Fig. 3. The results of various C, where $\varepsilon = 0.1$: (a) the S-MSE and (b) the number of support vectors.

by dual variables used as linear coefficients in SVM solution. Hence, a 'good' value for C can be chosen equal to the range of output values of training data [23]. However, the selection of parameter C also depends on the domain of application knowledge. In this study, ε is fixed to be 0.1 in one-time search method when varying the value of C from 2^{-5} to 2^{5}. At the same time, the number of support vectors is also recorded. Fig. 3(a) gives the results of various C where ε is fixed at 0.1. It shows that for every building, in terms of parameter C, there exists one lowest S-MSE point. It is also observed that for Buildings A–C, S-MSE first decreases slightly with parameter C increasing, and then increases after the optimum point. For Building D, S-MSE increases largely with increase of parameter C. Theoretically, a small value of C will under-fit the training data because the weight placed on the training data is too small thus resulting in large values of MSE on the test sets. However, when C is too large, SVM will over-fit the training set, which means that $\frac{1}{2}||\omega||^2$ will lose its meaning and the objective goes back to minimize the empirical risk only. In this study, obviously, the choice of parameter C should be the lowest point between 2^{-5} and 2^{5}, while it only varies slightly except for Building D. In Fig. 3(b), the number of support vectors increases slightly as C increases. The reason can be that when C gets larger, the optimization formulation (4) emphasized more on the empirical risk and makes the model fit the training data better at the cost of larger model complexity which means more support vector numbers are needed to determine the model. Moreover, referring to Eq. (10), larger C means larger range of value of support vectors, thus, more data points can be selected as the support vectors in the optimization formulation.

3.3.3.2. Selection of parameter ε. Parameter ε controls the width of the ε-insensitive zone, according to Fig. 1, which is normally used to fit the training data. Generally, the larger the ε, the fewer number of support vectors and thus the sparser the representation of the solution [12]. Furthermore, if the ε is too large, it can deteriorate the accuracy on the training data. As shown in Fig. 4(a), the MSE firstly remains constant, and then suddenly goes down during 2^{-2} to 2^{-3}, finally it slightly goes up. However, the number of support vectors shown in Fig. 4(b) decrease largely when the ε increases and eventually, they reach zero. This indicates that ε does not affect the performance of SVM much, while the number of support vectors shows a decreasing function of ε.

3.3.3.3. Stepwise search. The above method is simplest one in the literature to find out appropriate C and ε. Actually, there are several other methods developed to identify the best C and ε, among which, grid-search is frequently used, most complex and reliable one. In 'grid-search', all pairs of (C, ε) are tried and the one with the best performance is picked up. Chih-Wei Hsu [24] pointed out that exponentially growing sequences tries of C and ε was a practical way to identify good parameters (for example, defining $C = 2^{-5}$, 2^{-3}, …, 2^{5}, $\varepsilon = 2^{-13}$, …, 2^{-1}). However, the efficiency of the grid-search is low because it computes the performance at all pairs of C and ε to get the performance surface. Here, a more efficient way called stepwise method is developed to quickly search the acme of the performance surface. In the stepwise method, one-time search was first conducted to get MSE_1. Then, the same selection process is conducted again on parameter C (fixing the first result of ε) and ε (fixing the second result of C), to get lowest MSE_2. The one-time search

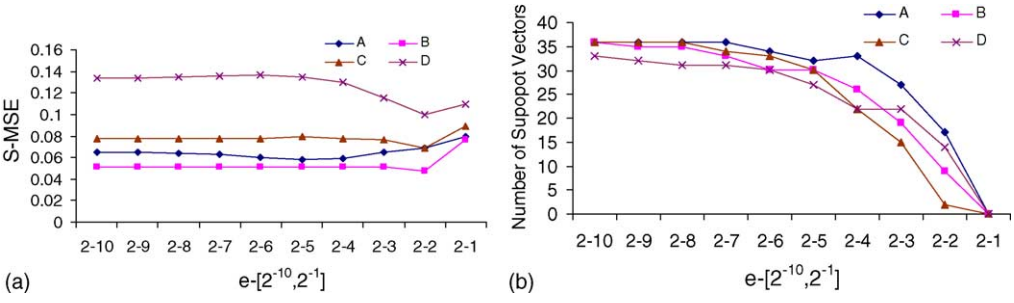

Fig. 4. The results of various ε where C is the result of selection: (a) the S-MSE and (b) the number of support vectors.

Table 3
Results of SVM on the test data

Building	C	ε	nSV	MSE	%Error	CV (%)
A	2^{3}	2^{-5}	32	0.73	−2.72	2.69
B	2^{-1}	2^{-2}	9	0.63	3.44	2.39
C	2^{0}	2^{-2}	2	0.17	0.68	1.28
D	2^{-5}	2^{-2}	14	0.14	−1.89	0.99

continues until $\mathrm{MSE}_n - \mathrm{MSE}_{n-1} < 0.00001$, and the training stopped. This stepwise search conserves the pairs trying character of grid-search and more accurate than one-time search, which is conducted only once on every parameter. Finally, after the best (C, ε) is found, the whole training set is trained again to generate the final regressor. For example, for Building A, ε is firstly found out and fixed at 0.03125, and then, parameter C, whose first result equals to 6, is searched again on the range from 2^{-5} to 2^{5}. Finally, C is chosen to be 8 and ε is 0.03125 with $\Delta\mathrm{MSE} = 0$. Thus, parameter C and ε; are eventually identified. Similarly, other three buildings are executed in the same way to find out best C and ε.

3.4. Results of SVM for prediction of landlord energy consumption

After parameters C and ε are selected, the final SVMs are confirmed. The summary of results of SVM is shown in Table 3. Table 3 shows that Building A has the highest MSE of 0.73, while Building D has the lowest MSE of 0.14. The number of support vectors is different from each building as well as parameter C and ε. In addition, all CVs, which represent the variances from the true value, are very small and less than 5%. This indicates all SVM models can be considered as excellent models according to Reddy et al. [19]. Comparing with other studies conducted using other methods such as neural networks (NN) and genetic programming (GP) on the building load research based on hourly or daily data, which are 1993 ASHRAE competition demonstrated the best CV of 10.36% (NN) on the whole building energy consumption, Kreider and his co-workers [3] found the best CV of 4.7% (NN) on chilled water and Chen et al. (2003) [26] found the best CV of 14.7% (GP) on the HVAC load, SVM in this study shows better results in terms of CV. The highest CV of 2.89% appears in Building A, while the lowest CV of 0.99% appears in Building D. Furthermore, the %error which shows the absolute difference between predicted value and real value is also small. The best %error appears in Building C, which is only 0.68%.

Fig. 5 shows the graphical results of predictions for four commercial buildings. Obviously, Building C shows the best result of prediction. Because of low CV in Building C, the predicted values are almost the same as real values. In addition, all four predicted value curves tracked the variation of real values correctly. It indicates that such kind of method can be applied in tracking the building energy use for diagnosing whether the systems are working orderly or not.

The performances of residues of four buildings are shown in Fig. 6. In ordinary least square (OLS) regression, the undesirable performance of residues is always a problem. In this study, it seems the same problem occurred again, which shows that the distributions of residues are not constant. However, since it appears randomly, the prediction model can be considered to be consistent.

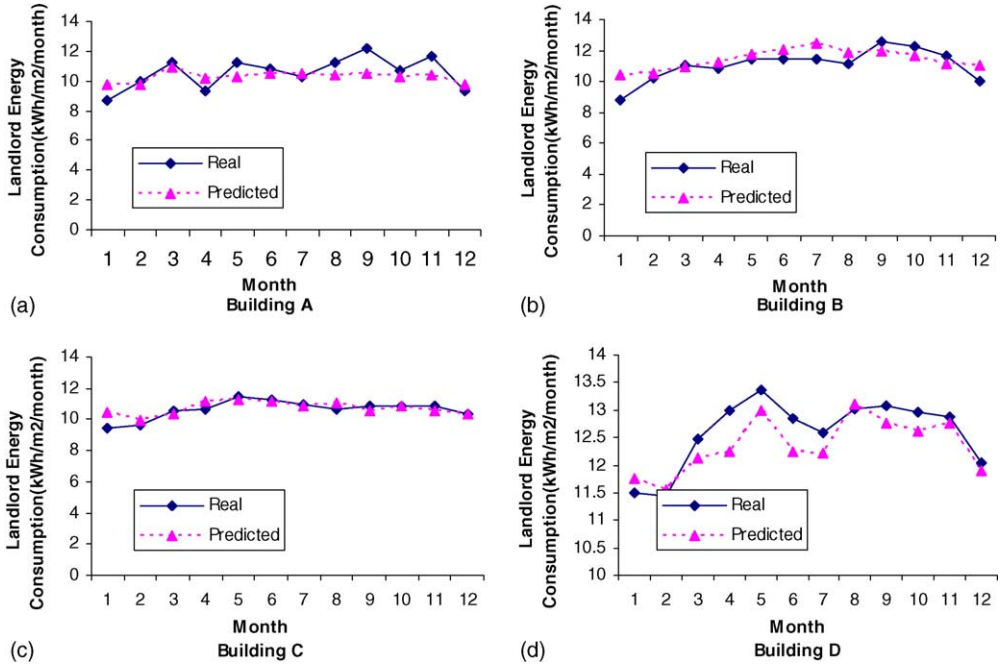

Fig. 5. Results of SVM for prediction of landlord energy consumption.

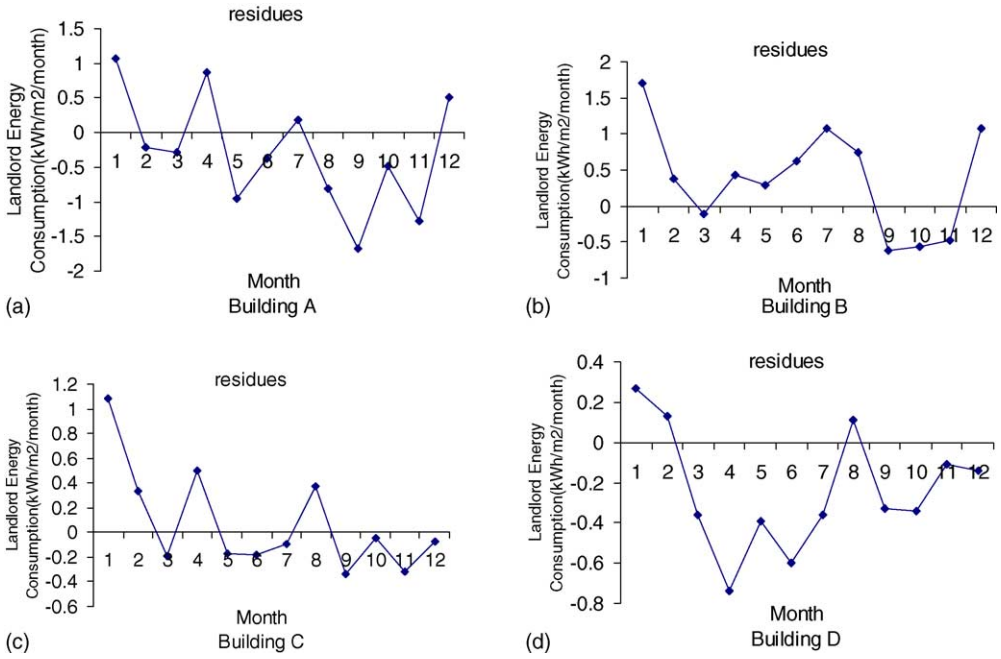

Fig. 6. Residues of estimated landlord energy consumption.

4. Conclusion

SVM is used to predict landlord energy consumption in this study. This is the first time that SVM is applied in the research of building load estimation. The performance of SVM, in terms of CV and MSE, is better than other related research using neural networks and genetic programming. The reasons can be as follows:

(1) Small data pool, the whole data profile consisting of only 4 years' monthly utility bills may include less scattered or abnormal data. However, other previous researches were based on the hourly or daily data, which covers thousands of data.
(2) Structural risk minimization (SRM) principle, which is the most outstanding feature of SVM, is implemented to minimize the upper bound of the generalization error rather than the training error, which is applied in NN. The detailed introduction of SRM can be referred to [14].
(3) There are fewer free parameters to optimize compared to neural network and genetic programming. As investigated in this study, only parameter C and ε are important parameters to the prediction models. However, for the neural networks, there are lots of free parameters needed to adjust such as number of neurons in the hidden layers, the learning rate, number of epochs, the stop criteria and the transfer functions. Furthermore, NN can never reach a global solution. However, the solution of SVM is unique and optimal because SVM is like solving a linearly constrained quadratic programming.

Furthermore, a stepwise search is developed in this study, which is more reliable yet simple. The final results demonstrates that SVM is feasible and applicable in prediction of monthly landlord utility bills in the tropical region. Moreover, the application of this methodology is not limited to only the tropical region based on its strongly theoretical background and regression characters. Since SVM presents many advantages in prediction, future research will focus on short-term load data exploration and prediction in commercial buildings.

References

[1] M. Fels, Special issue devoted to measuring energy savings, the Princeton scorekeeping method (PRISM), Energy and Buildings 9 (1–2) (1986).
[2] J.K. Kissock, A methodology to measure retrofit energy savings in commercial buildings, Doctoral Dissertation, Department of Mechanical Engineering, Texas A&M University, 1993.
[3] M. Krarti, J. Kreider, D. Cohen, P. Curtiss, Prediction of energy saving for building retrofits using neural networks, ASME Journal of Solar Energy Engineering 120 (3) (1998).
[4] A. Dhar, T.A. Reddy, D.E. Claridge, A. Fourier, Series model to predict hourly heating and cooling energy use in commercial buildings with outdoor temperature as the only weather variable, Journal of Solar Energy Engineering 121 (1999) 47–53.
[5] B. Dong, S.E. Lee, M.H. Sapar, A holistic utility bill analysis method for baselining whole commercial building energy consumption in Singapore, Energy and buildings, in press, 2004.
[6] B. Dong, S.E. Lee, M.H. Sapar, H.S. Sun, Applying neural networks to model monthly energy consumption of commercial buildings in Singapore, in: Proceedings of International Conference on Control, Automation and System, Thailand, 2004.
[7] B. Scholkopf, C. Burges, V. Vapnik, Extracting support data for a given task, in: Proceedings of First International Conference on

Knowledge Discovery and Data Mining, Menlo Park, CA, AAAI, 1995.
[8] V.N. Vapnik, S.E. Golowich, A.J. Smola, Support vector method for function approximation, regression estimation and signal processing, Advanced Neural Information Processing System 9 (1996) 281–287.
[9] R. Muller, J.A. Smola, B. Scholkopf, Prediction time series with support vector machine, in: Proceedings of International Conference on Artificial Neural Networks, 1997, pp. , p. 999.
[10] S. Mukherjee, E. Osuma, F. Girosi, Nonlinear prediction of chaotic time series using support vector machine, in: Proceedings of IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing VII (NNSP'97), Amelia Island, FL, 1997, pp. 511–520.
[11] R. Doss, Z. Obradovic, Regime signalling techniques for nonstationary time series forecasting, Neurovest Journal 4 (5) (1998) 7–15.
[12] L.J. Cao, Francis, E.H. Tay, Support vector machine with adaptive parameters in financial time series forecasting, IEEE Transactions on Neural Networks 14 (6) (2003).
[13] V. Cherkassky, Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, Neural Networks 17 (2004) 113–126.
[14] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[15] W.S. Sarle, 1997. Neural Network FAQ. Periodic posting to the Usenet news group comp.ai.neural-nets available: ftp://ftp.sas.com/pub/neural/FAQ.html.
[16] S.E. Lee, Energy efficiency of office buildings in Singapore, in: BCA Seminar on Energy Efficiency in Building Design, April 18, 2001.
[17] R. Sullivan, S. Nozari, R. Johnson, S. Selkowitz, Commercial building energy performance analysis using multiple regression, ASHRAE Transactions 91 (1985) 337–353.
[18] M.F. Fels, K.M. Keating, Measurement of energy savings from demand-side management programs in U.S. electric utilities, Annual Review Energy Environment 18 (1993) 57–88.
[19] T.A. Reddy, N.F. Saman, D.E. Claridge, J.S. Haberi, W.D. Turner, A. Chalifoux, Baselining methodology for facility-level monthly energy use—part 1: theoretical aspects, ASHRAE Transactions 103 (1997) 2.
[20] S.S. Keerthi, C.-J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, Neural Computation 15 (7) (2001) 1667–1689.
[21] H.-T. Lin, C.-J. Lin, 2003. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods, Technical report, Department of Computer Science and Information Engineering, National Taiwan University.
[22] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines. 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[23] D. Mattera, S. Haykin, Support vector machines for dynamic reconstruction of a chaotic system, in: B. Scholkopf, J. Burges, A. Smola (Eds.), Advances in Kernel Methods: Support Vector Machine, MIT Press, Cambridge, MA, 1999.
[24] C.-W. Hsu, C.-C. Chang, C.-J. Lin, 2003. A Practical Guide to Support Vector Classification, Technical report, Department of Computer Science and Information Engineering, National Taiwan University.
[25] S. Katipamula, T.A. Reddy, D.E. Claridge, Multivariate regression modeling, Journal of Solar Energy Engineering 120 (1998) 177–184.
[26] Z.Q. Chen, R.M. Nelson, D.A. Ashlock, Comparison of methods for predicting monthly post-retrofit energy use in buildings, ASHRAE Transactions 109 (2003) 449–459.
[27] N. Draper, H. Smith, Applied Regression Analysis, second ed. John Wiley and Sons, New York, 1981.