Electric Load Forecasting Based on Locally Weighted Support Vector Regression

Ehab E. Elattar, *Student Member, IEEE*, John (Yannis) Goulermas, *Member, IEEE*, and Q. H. Wu, *Senior Member, IEEE*

***Abstract*—The forecasting of electricity demand has become one of the major research fields in electrical engineering. Accurately estimated forecasts are essential part of an efficient power system planning and operation. In this paper, a modified version of the support vector regression (SVR) is presented to solve the load forecasting problem. The proposed model is derived by modifying the risk function of the SVR algorithm with the use of locally weighted regression (LWR) while keeping the regularization term in its original form. In addition, the weighted distance algorithm based on the Mahalanobis distance for optimizing the weighting function's bandwidth is proposed to improve the accuracy of the algorithm. The performance of the new model is evaluated with two real-world datasets, and compared with the local SVR and some published models using the same datasets. The results show that the proposed model exhibits superior performance compare to that of LWR, local SVR, and other published models.**

***Index Terms*—Load forecasting, locally weighted regression (LWR), locally weighted support vector regression (LWSVR), support vector regression (SVR), time series reconstruction, weighted distance.**

I. Introduction

LOAD forecasting has always been the essential part of an efficient power system planning and operation. It is always defined as basically the science or art of predicting the future load on a given system for a specified period of time ahead. Operation decisions in power systems, such as unit commitment, reducing spinning reserve, economic dispatch, automatic generation control, reliability analysis, maintenance scheduling, and energy commercialization, depend on the future behavior of loads. Therefore, accurate load forecasting helps the electric utility to make these operation decisions properly. The time series of power load is affected by many factors such as economic, temperature, etc., thus making it difficult to be accurately predicted.

Several short-term load forecasting (STLF) methods including traditional and artificial intelligence-based methods have been proposed during the last four decades. The relationship between electric load and its exogenous factors is complex and nonlinear, making it quite difficult to be modeled through traditional techniques such as linear or multiple regression [1], autoregressive moving average (ARMA), exponential smoothing methods [2], Kalman-filter-based methods [3], etc. On the other hand, various artificial intelligence techniques were used for STLF; among these methods, artificial neural networks (ANNs) have received the largest share of attention. The ANNs that have been successfully used for STLF are based on multilayered perceptrons [4]. The neural fuzzy network, which has a good performance in time series prediction [5], [6], has also been used for load forecasting. Its application in hourly load forecasting for 24 h ahead [7] has shown a better accuracy than that of the backpropagation neural network. Radial basis functions (RBFs) [8] have been also used for day-ahead load forecasting, giving better results than that of the conventional neural networks. In addition, the Bayesian neural network approach has been applied to STLF with success [9].

Support vector machine (SVM), which is proposed by Vapnik and coworkers [10], is a novel powerful machine learning method based on statistical learning theory (SLT). SVM replaces the empirical risk minimization (ERM) principle, which is generally employed in traditional ANN, by structural risk minimization (SRM) principle. The most important concept of SRM is the application of minimizing an upper bound to the generalization error instead of minimizing the training error. On the basis of this principle, SVM will be equivalent to solving a linear constrained quadratic programming problem, so that the solution of SVM is always unique and globally optimal. Originally, SVM has been developed for solving the classification problems and achieved good performances [11]–[13]. With the introduction of Vapnik's ε-insensitive loss function, SVM has been extended to solve the regression problems called support vector regression (SVR) [14]. Recently, SVR has been applied to various applications with excellent performances [15], [16]. Least-square SVM [17] and weighted least-square SVM [18] are also effective algorithms. The SVR has shown a high accuracy achieved when applied to solve the STLF problem [19], [20].

Because of the complexity of the historical load data and the uncertainty of the influencing factors, such as weather, economical, and random factors, the time series reconstruction technique can be applied to the power load forecasting. The correlation dimension method and the mutual information method are used to choose the optimal embedding dimension and time delay constant for the power load data. Our previous works [15], [21], [22] have shown that the local prediction methods based on phase

Manuscript received June 17, 2009; revised October 30, 2009; accepted December 31, 2009. Date of publication February 5, 2010; date of current version June 16, 2010. The work of E. E. Elattar was supported by the Egyptian Government under The University of Liverpool Ph.D. scholarship. This paper was recommended by Associate Editor Y. Jin.

E. E. Elattar is with the Department of Electrical Engineering and Electronics, The University of Liverpool, Liverpool L69 3GJ, U.K., and also with the Department of Electrical Engineering, Minufiya University, Shebin El-Kom 32511, Egypt (e-mail: elattar@liverpool.ac.uk).

J. Y. Goulermas and Q. H. Wu are with the Department of Electrical Engineering and Electronics, The University of Liverpool, Liverpool L69 3GJ, U.K. (e-mail: j.y.goulermas@liverpool.ac.uk; q.h.wu@liverpool.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCC.2010.2040176

reconstruction can provide generally better results than those obtained with global methods based on phase reconstruction.

Locally weighted regression (LWR) [23] is a kind of locally weighted learning methods. LWR forms a local model around a point of interest whereby only training data that are closest to that point will be used in handling each query, instead of using all training data [24]. After answering the query, the aforesaid local model is discarded. To answer a new query, a new local model is created, which means that every set of training and generation period is unique and independent of others. LWR is a method for estimating a regression surface through multivariate smoothing: the response variable is smoothed dynamically, as a function of the predictor variables [23]. LWR consists of developing a moving local model to a set of nearest neighbors.

In this paper, we proposed a new approach to a load forecasting by combining the SVR and LWR, which can be called as locally weighted support vector regression (LWSVR). LWSVR is an ameliorated SVR, which endows a weight factor to each train load datum. The weighting function's bandwidth plays an important role in local modeling. So, the weighted distance algorithm that uses the Mahalanobis distance is proposed to optimize this bandwidth. Two datasets have been used for evaluating the proposed model. The first dataset is related to a daily peak load forecasting competition (European Network of Excellence on Intelligent Technologies for Smart Adaptive Systems (EUNITE) competition), with load and temperature data from the Eastern Slovakian Electricity Corporation [25], while the second dataset corresponds to the load and temperature series, on an hourly basis, from a North American electric utility [26].

The paper is organized as follows. Section II summarizes the previous related work. Section III outlines the basics of time series reconstruction. Section IV describes the LWSVR algorithm. Section V presents the weighted distance algorithm used to optimize the bandwidth. Experimental results and comparisons on load forecasting problem are presented in Section VI. Finally, Section VII concludes the paper.

II. Related Work

In global predictors, a prediction model is trained based on the entire data history and used to predict the load at a specific time with a fixed data window. Such an approach has the disadvantage that if new information is taken into consideration, all parameters of the model may need to be updated, and also a lengthy parameter reestimation stage is required. Another disadvantage is concerned with its disability of capitalizing the historical information of the time series directly, as only the current window is used for prediction at a specific time.

To overcome the drawbacks of the global predictors, the local predictors can be used [15]. They overcome the drawbacks of global predictors by utilizing part of the relevant history directly in the prediction model. Specifically, only the set of points of the reconstructed space that are close enough to the point under prediction is used to fit the local function. Another advantage of the local predictors is that the training set for each point on the reconstructed trajectory is much smaller than the global predictors, which require the use of all available training examples, and they can save memory space through decomposing the prediction problem to several smaller ones.

In the last few years, the local predictor approach has interested many researchers to solve the nonlinear time series prediction problem such as [27] and [28]. McNames *et al.* [29] introduced the local averaging model for time series prediction. This method, which can be used with smaller neighborhoods, is more stable and often more accurate than local linear model for very short dataset. Therefore, this model was used in [29] to generate the winning entry of the Katholieke Universiteit Leuven time series prediction competition. Lau *et al.* [15] combined the strength of SVR and local predictor. The proposed algorithm, local SVR, gave a better prediction results than other local models when it is applied to nonlinear time series prediction.

In our previous work [21], [22], we combined a proven powerful regression algorithm (SVR) with a local prediction framework to solve the STLF problem. In this approach, the embedding dimension and the time delay constant for the power load data are computed first, and then the continuous power load data are used for the phase space reconstruction. In addition, the neighboring points are presented by Euclidian distance. According to these neighboring points, the local model is set up.

In all aforementioned techniques, the regularization parameter of SVR is constant, so that all training data contribute to the accuracy of the model to the same extent. However, in many cases, the effects of the training points are different where some training points are more important than others. Therefore, the model should have higher accuracy for these training points.

To achieve this goal, some trials are made to modify the standard SVR by weighting the SVR's regularization parameter. Tay *et al.* [30] introduced a modified SVM for financial time series forecasting, which is called C-ascending SVM. In this method, each data are weighted using one of the two designed weighting functions. They are linear weight function and exponential weight function. Then, the recent historical data points have larger weights than the distant historical data. Lee *et al.* [31] proposed the weighted SVM for quality estimation in the polymerization process. The proposed method combines the SVM and LWR. Each data are weighted according to its distance to the current prediction point. Hu *et al.* [32] introduced the weighted SVM-based fuzzy C-mean clustering algorithm to solve STLF problem. In this approach, the training samples are clustered into several subsets with consideration of homogenous characteristics. In addition, according to the time, each data are weighted. The older data points have smaller weights than the new ones. Unfortunately, the earlier trials were used as global predictors except [31].

The study presented here extends our previous work [21], [22] by modifying the risk function of the standard SVR with the use of LWR while keeping the regularization term in its original form. In addition, to optimize the weighting function's bandwidth, the weighted distance algorithm that uses the Mahalanobis distance is presented.

Our approach is different from the previous works. First, we reconstruct the phase space of time series using the embedding dimension (d) and the time delay constant (m). Here, the correlation dimension method and the mutual information method

are used to calculate d and m, respectively. Second, the Euclidian distance is used to find the neighboring points for each query point. Then, each point in the neighborhood is weighted according to its distance from the query point to calculate the new regularization parameter of SVR. Moreover, these neighboring points are used only to train the prediction model instead of using all available training points, so that the drawbacks of global predictors can be overcome. Finally, the weighted distance algorithm is used to optimize the bandwidth of the weighting function, so that we can overcome the disadvantage of using this bandwidth as a fixed value.

III. Time Series Reconstruction

A commonly used tool for the study of complex time series and dynamical systems is the phase space reconstruction technique that stems from the embedding theorem developed by Takens [33] and Sauer *et al.* [34]. Suppose we have an univariate time series $x(t)$ for $t = 1, 2, \ldots, N$, where N is the length of the dataset, and regarding the embedding theorem, $x(t)$ can be extending to a vector $z(t)$ in a d-dimensional space as follows:

$$z(t) = [x(t), x(t-m), x(t-2m), \ldots, x(t-(d-1)m)] \tag{1}$$

where d is called the embedding dimension of the system and m is the delay constant. The time delay constant and the embedding dimension are two important parameters for reconstructing phase space. In this paper, the correlation dimension method and the mutual information method are used to compute d and m, respectively.

For the multivariate time series, assuming that there are n time series, they are $\{x_i(t)\}$ $(i = 1, 2, \ldots, n)$. According to the embedding theorem developed by Takens [33] and Sauer *et al.* [34], the reconstructed vector of multivariate time series in the phase space could be denoted as [35]

$$z_i(t) = [x_i(t), x_i(t-m_i), x_i(t-2m_i), \ldots, x_i(t-(d_i-1)m_i)] \tag{2}$$

where $i = 1, 2, \ldots, n$; $t = 1, 2, \ldots, L$. The variables d_i and m_i are the selected embedding dimension and time delay constant of the ith time series, respectively. L is the length of the embedded points generated in the phase space, which can be computed using the formula $L = N - \max_{i=1,\ldots,n}[(d_i - 1)m_i]$. $z(t)$ is now an $L \times D$ matrix, where $D = \sum_{i=1}^{n} d_i$. The details of how to choose the proper values of d and m using the correlation dimension method and mutual information method have been reported in [15].

IV. Locally Weighted Support Vector Regression

A. Support Vector Regression

There are two key features in the implementation of SVR. They are quadratic programming and kernel functions. By solving a quadratic programming problem with linear equality and inequality constraints, the SVR's parameters can be obtained. The flexibility of kernel functions allows the technique to search a wide range of the solution space [36].

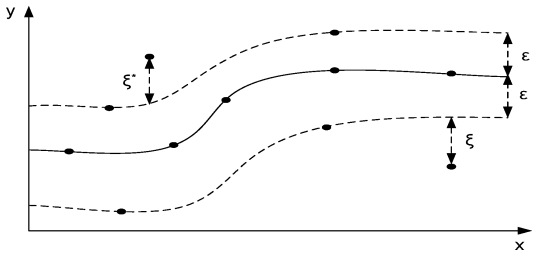

Fig. 1. ε-insensitive tube for SVR.

Suppose there is a set of training data $\{x_i, y_i\}_{i=1}^{N}$, where each $x_i \in \Re^n$ denotes the input space of the sample and has a corresponding target value $y_i \in \Re$ for $i = 1, \ldots, N$, with N corresponding to the size of the training data. The SVR's basic idea is to find a nonlinear map from input space to output space and map the input data to a higher dimensional feature space through this map. Then the following estimate function is used to make linear regression in that feature space [14] as

$$f(x) = \langle w, \phi(x) \rangle + b \tag{3}$$

where $\phi(x)$ denotes the high-dimensional feature space, which is nonlinearly mapped from the input space, w contains the coefficients that have to be estimated from the data, and b is a real constant. The objective is to minimize the following risk function [10]:

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$
$$\text{subject to} \begin{cases} y_i - \langle w, \phi(x_i) \rangle + b \le \varepsilon + \xi_i^* \\ \langle w, \phi(x_i) \rangle + b - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{4}$$

where x_i is mapped to higher dimensional space by the function ϕ, ξ_i is the lower training error (ξ_i^* is the upper) subject to the ε-insensitive tube $|y - \langle w, \phi(x) \rangle + b)| \le \varepsilon$, the term $1/2\|w\|^2$ is called the regularization term, and C is the regularization constant that determines the tradeoff between the flatness of f and its accuracy in capturing the training data.

The constraints of (4) imply that the most of the data x_i are placed inside the tube ε. If x_i is outside the tube, there is an error ξ_i or ξ_i^* that we tend to minimize in the objective function. This can be seen in Fig. 1. SVR avoids underfitting and overfitting of the training data by minimizing the regularization term $1/2\|w\|^2$ as well as the training error $C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$.

Introducing Lagrange multipliers α_i and α_i^* with $\alpha_i \alpha_i^* = 0$ and $\alpha_i, \alpha_i^* \ge 0$ for $i = 1, \ldots, N$, and according to the Karush–Kuhn–Tucker optimality conditions [14], the SVR training procedure amounts to solve the convex quadratic problem

$$\min_{\alpha,\alpha^*} \begin{cases} \dfrac{1}{2}\displaystyle\sum_{i,j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)Q(x_i, x_j) \\ \\ +\varepsilon\displaystyle\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \end{cases}$$

$$\text{subject to} \begin{cases} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C \end{cases} \tag{5}$$

where $Q(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ is the kernel function that is the inner product of the points $\phi(x_i)$ and $\phi(x_j)$ mapped into feature space. Using the kernels, all necessary computations can be undertaken directly in the input space, without calculating the explicit map $\phi(x)$. The typical examples of kernel function are linear, polynomial, Gaussian, etc. [14]. Choosing a good one for SVR algorithm is a problem itself. Therefore, different types like linear, polynomial, and Gaussian are tested before choosing the best one. We found that the Gaussian kernel gives better results than other types in the range of 8.91%–10.43%. So, in all test cases in this paper, we employ the commonly used Gaussian kernel defined as

$$Q(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right). \tag{6}$$

Finally, the regression output takes the following form:

$$\hat{f}(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)Q(x_i, x) + b. \tag{7}$$

B. Local Support Vector Regression

Local prediction is predicting the future based only on a subset of training data to find a local $\hat{f}$, which approximate the function f locally in the reconstructed embedded space. This subset (neighborhoods) is chosen on the basis of the Euclidian distance between the testing data and the training data in the input space. In the local SVR model, the dynamics of time series systems can be captured step by step locally in the phase space and the drawbacks of global methods can be overcome.

The local SVR model can be summarized as follows [21], [22]. First, reconstruct the time series using d and m as described in Section III. For each query vector q, the K nearest neighbors $\{z_q^1, z_q^2, \ldots, z_q^K\}$ among the training inputs are chosen using the Euclidian distance as the distance metric between the q and each z in the reconstructed time series. Using these K nearest neighbors, train the SVR to obtain support vector and corresponding coefficients. Finally, the output $\hat{f}(q)$ can be computed by (7).

C. Locally Weighted Regression

LWR is derived from standard linear regression. This algorithm fits a surface to "local" points using distance-weighted regression. LWR is based on the assumption that the neighboring values of the predictor variables are the best indicators of the response variable in that range of predictor values [23].

To estimate the value of the function $\hat{f}(x)$ at any value of x in the d-dimensional space, the K (neighborhood size) data points whose x_i values are closest to x are used ($1 < K \ll N$). Each point in the neighborhood is weighted according to its distance from x. The points that are close to x have large weights, and the points far from x have small weights [31].

Many weighting functions are proposed by the researchers [23], [37]. Out of these weighting functions, Gaussian kernel, tricube kernel, and quadratic kernel are the most popular [24]. The widely used weighting function is the Gaussian kernel weighting function [37], which can be defined as follows:

$$W(d_E) = e^{-(\frac{d_E}{h})^2} \tag{8}$$

where h is the bandwidth parameter that plays an important role in local modeling. From (8), the weight of the data point (x_i, y_i) is then

$$W_i = W\big(\sqrt{(x - x_i)^T(x - x_i)}\big). \tag{9}$$

Thus, W_i has its maximum value when x_i is closest to x, and decreases as x_i increases in distance from x.

D. Locally Weighted Support Vector Regression

The presented model is derived by modifying the risk function of the standard SVR with the use of LWR. The constant C in (4) is the regularization constant that determines the tradeoff between the flatness of f and its accuracy in capturing the training data. When C is A constant, all training data contribute to the accuracy of the model to the same extent [31]. In the load forecasting problem, it is common that some training points are more important than others [21], [22]. Therefore, the model should have higher accuracy for the training input data that are closer to the new input point for prediction.

To achieve this goal, C is computed as a function of the distance between input data points and the concept of LWR is used. In the presented approach, the modified risk function can be formulated as follows:

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} C_i(\xi_i + \xi_i^*) \tag{10}$$

and

$$C_i = W_i \times C \tag{11}$$

where W_i is the weight function obtained from (9). Replacing the constant C in (5) using (11), the dual problem's constraints can be written as

$$\text{subject to} \begin{cases} \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i, \alpha_i^* \le C_i. \end{cases}$$

By solving this problem, the regression output can be obtained using (7).

V. Weighted Distance Algorithm For Optimizing The Bandwidth

A bandwidth parameter h defines the scale or range over which generalization is performed. This is a very important parameter that plays an important role in local modeling. If h is infinite, then the local modeling becomes global. On the other hand, if h is too small, then it is possible that we will not have

adequate number of data points in the neighborhood for a good prediction.

There are several ways to use this parameter like fixed bandwidth selection where h is constant, nearest neighbor bandwidth selection where h is set to be the distance between the query point and the Kth nearest point, global bandwidth selection where h is calculated globally by an optimization process [24], etc. In fixed bandwidth selection method, h is chosen as a constant value; therefore, the training data with constant size and shape are used. However, it is the easiest way to adjust the radius of the weighting function, and its performance is unsatisfactory for nonlinear system as the density and distribution of data points are unlikely to be identical at every place of the dataset [38]. In this paper, we used the weighted distance algorithm that uses the Mahalanobis distance for optimizing the bandwidth (h) to improve the accuracy of our proposed method (i.e., the improvement in the accuracy using this method over the nearest neighbor bandwidth selection method is in the range of 2.50%–4.00%).

The Mahalanobis distance is based on correlation between variables by which different patterns can be identified and analyzed. With this measure, the problem of scale and correlation inherent in Euclidean distance is no longer an issue. In the Euclidean distance, the set of points that have equal distance from a given location is a sphere. The Mahalanobis distance stretches this sphere correct for the respective scales of the different variables and account for correlation among variables.

The standard Mahalanobis distance can be defined as

$$\mathrm{MD}(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \tag{12}$$

where x is a vector of data, μ is a mean, and S^{-1} is inverse covariance matrix.

Defining the Mahalanobis distance between the query point x_q and data point x as $\mathrm{MD}_q = \sqrt{(x-x_q)^T S^{-1}(x-x_q)}$, where x belongs to the K nearest neighbors of the query point x_q and S^{-1} is computed after removing the mean from each column, the bandwidth h_q is the function of MD_q given as

$$h_q = \Theta(\mathrm{MD}_q) \tag{13}$$

where $\mathrm{MD}_{\min} \le \mathrm{MD}_q \le \mathrm{MD}_{\max}$ and $\mathrm{MD}_{\min}$ is the distance between x_q and the closest neighbor, while $\mathrm{MD}_{\max}$ is the distance between x_q and the farthest neighbor.

According to the LWR method, the point corresponding to $\mathrm{MD}_q = \mathrm{MD}_{\min}$ is most important that is $h_{\max} = \Theta(\mathrm{MD}_{\min}) = 1$, while the point corresponding to $\mathrm{MD}_q = \mathrm{MD}_{\max}$ is the least important, i.e., $h_{\min} = \Theta(\mathrm{MD}_{\max}) = \delta$, where δ is a real constant. This constant is a low-sensitivity parameter. Therefore, after few trials, we fix it to 0.01, which gives the best results.

The bandwidth h_q can be selected as a function of MD_q as follows [38]:

$$h_q = \Theta(\mathrm{MD}_q) = a\left(\frac{1 - b\,\mathrm{MD}_q}{\mathrm{MD}_q}\right)^2 + c \tag{14}$$

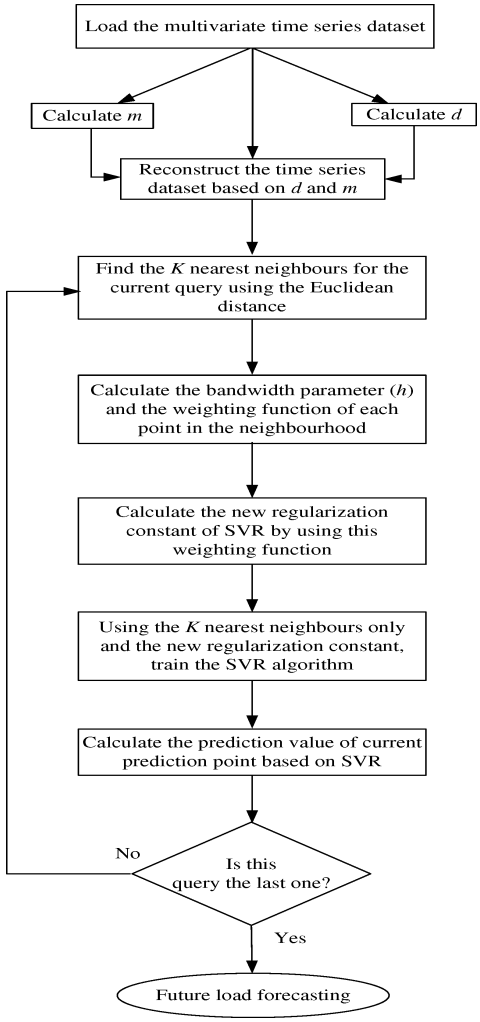

Fig. 2. Flowchart of the proposed model.

where a, b, and c are constants. By applying the boundary conditions, we can calculate these constants and get [38]

$$h_q = \Theta(\mathrm{MD}_q) = (1-\delta)\left(\frac{\mathrm{MD}_{\min}(\mathrm{MD}_{\max} - \mathrm{MD}_q)}{\mathrm{MD}_q(\mathrm{MD}_{\max} - \mathrm{MD}_{\min})}\right)^2 + \delta. \tag{15}$$

The Gaussian kernel weighting function that is used in this paper can be written as follows:

$$W(\mathrm{MD}_q) = e^{-\left(\frac{\mathrm{MD}_q}{(1-\delta)\left(\frac{\mathrm{MD}_{\min}(\mathrm{MD}_{\max}-\mathrm{MD}_q)}{\mathrm{MD}_q(\mathrm{MD}_{\max}-\mathrm{MD}_{\min})}\right)^2+\delta}\right)^2}. \tag{16}$$

Fig. 2 presents the computation procedure of the proposed model that can be divided to four main stages. The first stage reconstructs the multivariate time series using the embedding dimension and the time delay constant. The second stage finds the K closest vectors, or nearest neighbors, of the observed variables in the dataset for each query vector, and calculates the bandwidth parameter (h) and weighting function of each point in the neighborhood. Then it calculates the modified risk function of SVR. The third stage trains the SVR with modified risk function using only the K nearest neighbors. The fourth

TABLE I
PHASE RECONSTRUCTION PARAMETERS FOR EACH DATASET

Dataset	Load time series		Temperature time series		K
	d_1	m_1	d_2	m_2	
EUNITE competition	4	2	4	2	34
North American electric utility	4	5	3	9	35

stage evaluates the model using the query vector as the input to estimate the process output.

VI. EXPERIMENTAL RESULTS

A. Datasets

To evaluate the performance of the proposed LWSVR model, two different datasets are used. The first one is the data provided by the EUNITE network during the daily peak load competition [25], while the second one is the hourly load and temperature from North American electric utility [26].

For the first dataset, the organizer of the competition had provided the following data to the competitors: half hourly electricity load demand from January 1997 to December 1998, average daily temperature from 1995 to 1998, and holiday's information from 1997 to 1999. While for the second dataset, the hourly load and temperature from January 1985 to March 1991 are available.

Certain characteristics can be reported about the datasets before evaluating our proposed model. The load in both datasets has some seasonal patterns: the electricity demand in winter period is higher than the electricity demand in summer period. This implies the relation between electricity usage and weather conditions in different seasons. Also, the load has daily and weekly periodicity. Load demand in weekdays (monday–friday) is usually higher than that of weekend. In addition, electricity demand in sunday is a little lower than that on saturday in the first dataset, while it is a little higher than that on saturday in the second dataset.

B. Parameters

To implement a good model, there are some important parameters to choose. Choosing the proper values of d and m is a critical step in the algorithm. The correlation dimension method and the mutual information method are used to selecte d and m, respectively, and the optimal values of these parameters are shown in Table I. Using the obtained values of d and m, the multivariate time series can be reconstructed as described in Section III.

Also, choosing K is a very important step in order to establish the local prediction model. There are some methods used in literature to find this parameter such as cross validation [39] and bootstrap [40]. This parameter should be high for low-density datasets, while it should be low for high density ones. So, in this paper, we calculate K using the following method [21]:

$$K = \text{round}\left(\frac{\alpha}{N \times k_{\max} \times D_{\max}} \sum_{i=1}^{N} \sum_{k=1}^{k_{\max}} D_k(x_i)\right) \quad (17)$$

where N is the number of training points, $k_{\max}$ is the maximum number of nearest neighbors, $D_k(x_i)$ is the distance between each training point x and its nearest neighbors, while $D_{\max}$ is the maximum distance, $\frac{1}{N \times k_{\max}} \sum_{i=1}^{N} \sum_{k=1}^{k_{\max}} D_k(x_i)$ is the average distance around the points, which is inversely proportional to the local densities, and α is a constant. The two constants $k_{\max}$ and α are very low-sensitivity parameters. $k_{\max}$ can be chosen as a percentage of the number of training points (N) for efficiency, while α can be chosen as a percentage. In this paper, $k_{\max}$ and α are always fixed for all test cases at 30% of N and 75, respectively. Table I shows the value of K for each test case used in this paper.

There are some key parameters for SVR, which are C, ε, and σ in the Gaussian kernel function. The selection of these parameters is important to the generalization of the forecasting. Therefore, in order to get these parameters, we divided the training data into two subsets. One of them is used to train the model while the other is used to validate the model. Based on this partition, the suitable parameters are chosen using the following procedures [41].

1) Set initial values of C and ε. Then, adjust the value of σ till a minimum validation error is achieved.
2) Fix the value of ε and use the value of σ as calculated in the previous step. Then, adjust the value of C till a minimum validation error is achieved.
3) Use the values of σ and C as calculated in the previous two steps. Then, adjust the value of ε till a minimum validation error is achieved.

C. Numerical Results

For all performed experiments, we quantified the prediction performance with mean absolute error (MAE), mean absolute percentage error (MAPE), normalized mean square error (NMSE), and relative error percentage (REP). They can be defined as

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |A_i - F_i| \quad (18)$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|A_i - F_i|}{A_i} \times 100 \quad (19)$$

$$\text{NMSE} = \frac{1}{\Delta^2 N} \sum_{i=1}^{N} (A_i - F_i)^2 \quad (20)$$

$$\Delta^2 = \frac{1}{N-1} \sum_{i=1}^{N} (A_i - \overline{A})^2$$

$$\text{REP} = \sqrt{\frac{\sum_{i=1}^{N} (A_i - F_i)^2}{\sum_{i=1}^{N} A_i^2}} \times 100 \quad (21)$$

TABLE II
COMPARISON OF THE LWSVR MODEL AND OTHER MODELS USING THE DATASET OF EUNITE COMPETITION

Prediction model	MAE (MW)	MAPE	NMSE	REP
LWR	20.31	1.78	0.06	1.89
Local SVR	15.40	1.53	0.05	1.72
LWSVR	13.76	1.34	0.03	1.54

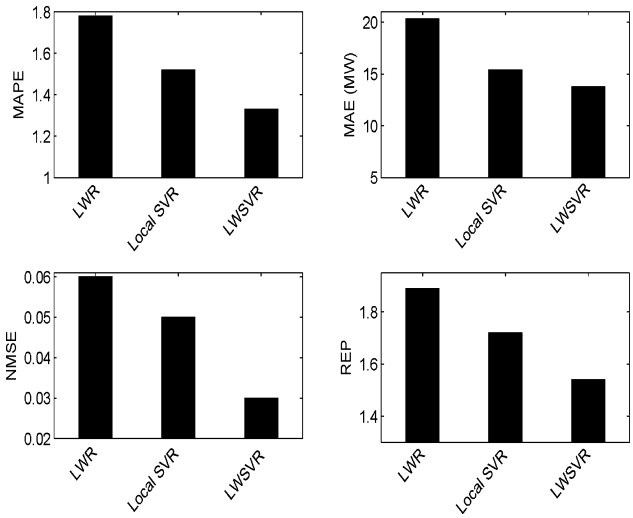

Fig. 3. Comparison of LWSVR model and other models using the dataset of EUNITE competition.

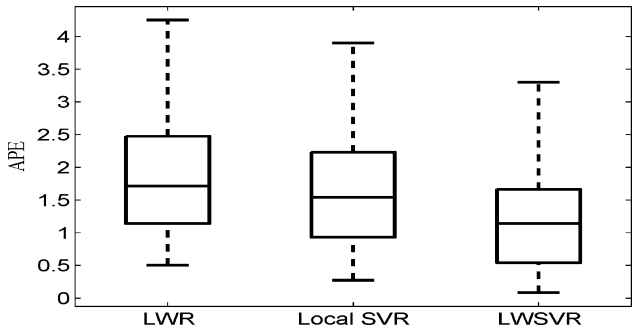

Fig. 4. APE results of January 1999 for each method.

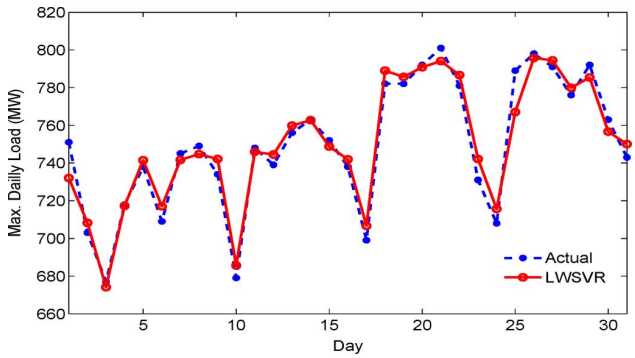

Fig. 5. Forecasted and actual maximum daily load in January 1999.

where A_i, F_i, and $\overline{A}$ are the actual value, the forecasted value, and the mean of the actual values, respectively, N is the testing dataset size, and i denotes the test instance index.

We designed two cases to evaluate the performance of the proposed model.

Case 1: In this case, we used the dataset provided in EUNITE competition. Our goal in this case as well as the goal of the competition is to forecast the maximum daily load for January 1999. To achieve this goal, the load and temperature information of winter data (from January to March and October to December in 1997 and 1998) are used as a training period.

The performance of the LWSVR model is compared with LWR and local SVR models. This comparison is shown in Table II and depicted in Fig. 3.

These results show that the LWSVR model outperforms LWR and local SVR models. It improves the MAPE over LWR and local SVR models by 24.72% and 12.42%, respectively. In addition, the absolute percentage errors (APEs) between the forecasted load and actual load values of January 1999 for each method are calculated. Fig. 4 presents box plots of the APE results for LWR, local SVR, and LWSVR methods. These results confirm the superiority of our proposed method over other methods. Moreover, the actual load and forecasted load values of the peak daily load of January 1999 are plotted in Fig. 5. The results of Fig. 5 show that our prediction values are very close to the actual values.

The Mann–Whitney U test [42] is carried to confirm the significance of forecasting accuracy of the LWSVR over LWR and local SVR methods. Mann–Whitney U test is an approach for assessing the significance of a difference in the central tendency of two data series. The test is conducted at the 0.05 significance levels in one-tail test. The P values for (LWSVR versus LWR) and (LWSVR versus Local SVR) are 0.0095 and 0.0147, respectively, which is less than 0.05. Therefore, the LWSVR model provides significantly better forecast results than LWR and Local SVR.

To further study the superiority of LWSVR over other published models, its performance is compared with some published models that employ the dataset of EUNITE competition. These models are:

1) model A [19]: SVM;
2) model B [43]: SVM optimized by genetic algorithm;
3) model C [44]: SVM -based input dimension reduction;
4) model D [45]: extended Bayesian training method.

All of these models are global models. During the EUNITE competition, the real temperature of January 1999 was not provided to the competitors, so that the load data were only used by some competitors like Chen *et al.* [19] to forecast the maximum daily load in January 1999. SVM was employed in [19] to generate the winning entry of the EUNITE competition using the load data from January to March and October to December in 1997 and 1998, without using the temperature information. Genetic algorithm was used in [43] to choose the optimal values of the SVM's parameters. A new model was proposed in [44] to load forecasting by establishing the feature selection model and using floating search method to find the feature subset. Then SVM can be used to forecast the load using a small sample of the data. Moreover, after the competition was closed, the real temperature of January 1999 was available. This encourage some

TABLE III
COMPARISON OF THE LWSVR MODEL AND OTHER MODELS USING THE DATASET OF EUNITE COMPETITION (MAPE)

Input data	Training period	MAPE					Improvement
		LWSVR	Model A	Model B	Model C	Model D	
Load only	Load data of January to March and October to December in 1997 and 1998	1.41	1.95	–	–	–	27.69%
	Load data for two years (1997 and 1998)	1.44	–	1.93	1.70	–	25.39% and 15.29%, respectively
Load and temperature	Load and temperature data for two years (1997 and 1998)	1.38	–	–	–	1.75	21.14%

TABLE IV
COMPARISON OF THE LWSVR MODEL AND OTHER MODELS USING THE DATASET OF NORTH AMERICAN ELECTRIC UTILITY

Prediction model	MAE (MW)	MAPE	NMSE	REP
Model D [45]	–	4.88	–	–
Model E [46]	–	4.73	–	–
LWR	139.11	4.71	0.13	4.87
Local SVR	121.84	4.08	0.11	4.47
LWSVR	101.02	3.62	0.08	3.98

researchers like Ferreira and da Silva [45] to use the temperature information to forecast the maximum daily load of January 1999. The extended Bayesian training method was used in [45] and gave better results than that of the competition's winner.

More details about the data used and the training period of each model can be found in Table III. Moreover, the Gaussian kernel function (6) is not only used in all our experiments, but also in [19], [43], and [44]. To compare our proposed model with models A, B, C, and D, we used the same experimental setup as used in each model. This comparison is shown in Table III. The results of Table III show that the LWSVR outperforms other models.

Case 2: In this case, we used the dataset of the North American electric utility. At this electric power utility, the daily forecasts were made at 8:00 A.M. Forecasts were produced for the entire next day, starting at midnight and through the following midnight, and hence, they were made from 16 to 40 h in the future. On Friday, the forecasts were produced for the entire weekend as well as Monday (16–88 h into the future). The objective is to forecast the hourly load, from 16 up to 40 h (steps) ahead for weekdays and from 16 up to 88 h (steps) ahead for weekends during the test period that goes from November 1990 to March 1991.

To compare the proposed LWSVR with some published models that employ the same dataset, we used the same experimental setup as used in [45] (model D) and [46] (model E), which use a multiple regression model called EGRV. That is, the hourly load and temperature data from the month to be forecasted and from two month earlier, along with the data corresponding to the same window in the previous year are used as a training period. First, we calculate the error of each day during the testing period. Then the average error of each day of the week (Monday to Sunday) during the testing period is calculated. Finally, the overall mean performance for the entire testing period for each model can be calculated. These results are summarized in Table IV and depicted in Fig. 6.

It can be seen from these results that the LWSVR model gives better performance than LWR, local SVR, model D [45], and model E [46]. It improves the performance (MAPE) over LWR, local SVR, model D [45], and model E [46] by 23.14%, 11.27%, 25.82%, and 23.47%, respectively.

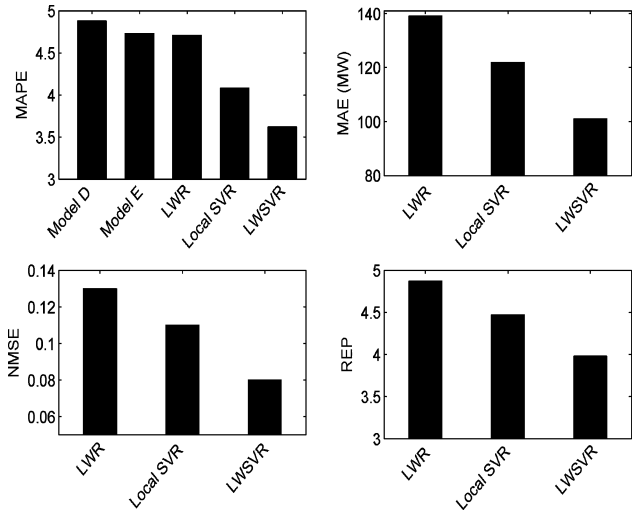

Fig. 6. Comparison of LWSVR model and other models using the dataset of North American electric utility.

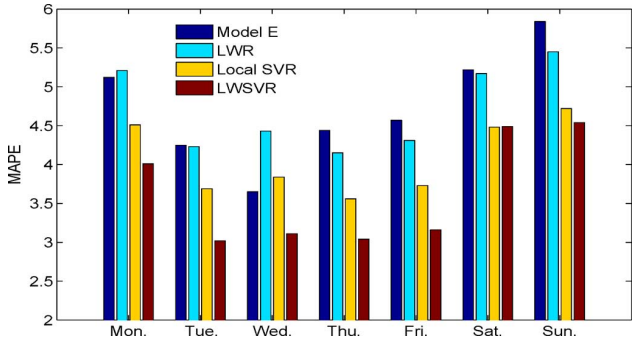

Fig. 7. Average prediction MAPE of every day of the week during the testing period.

Fig. 7 shows the average prediction MAPE of every day of the week (Monday–Sunday) during the testing period. These results confirm the superiority of the LWSVR model over other models.

In addition, the MAPE of the whole testing data for the 24 h is calculated. Fig. 8 presents box plots of these MAPE results for model E [46], LWR, local SVR, and LWSVR methods. From these results, we can note that the LWSVR model exhibits a better performance than LWR, local SVR and Model E [46]. Most

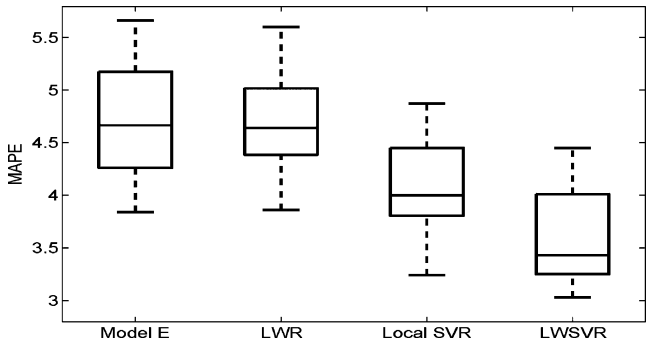

Fig. 8. Comparison of the LWSVR model and other models (MAPE of the 24 h).

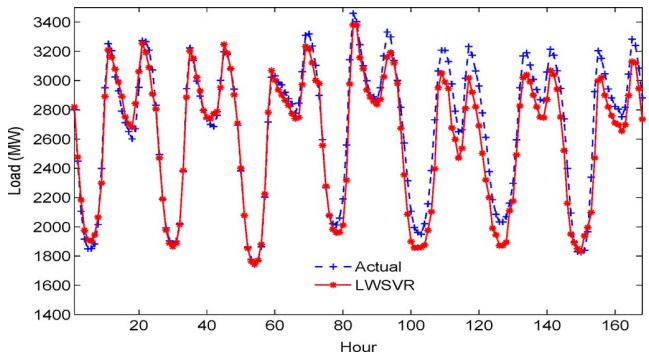

Fig. 9. Forecasted and actual hourly load from November 27, 1990 to December 3, 1990.

MAPE values of the proposed model are under 4, while in local SVR, half MAPE values are under 4. Also, we can note that most MAPE values of Model E [46] are over 4. The Mann–Whitney U test [42] is conducted at the 0.05 significance levels in one-tailed test to confirm the significance of forecasting accuracy of the LWSVR over Model E [46], LWR, and local SVR methods. The P values for (LWSVR versus Model E [46]), (LWSVR versus LWR), and (LWSVR versus Local SVR) are 0.0025, 0.0031, and 0.0063, respectively, which is less than 0.05. These results show that the LWSVR model provides significantly better forecast results than model E [46], LWR, and local SVR.

Fig. 9 presents one example for the dataset of the North American electric utility. It shows the forecasted hourly load versus the actual load of the period from November 27, 1990 to December 3, 1990. These results show that our prediction values are very close to the actual values.

VII. Conclusion

In this paper, we have proposed a new approach that can be used to solve the load forecasting problem. The approach combines the SVR and LWR, and employs the weighted distance algorithm that uses the Mahalanobis distance to optimize the weighting function's bandwidth. In the proposed model, the phase space is reconstructed based on multivariate time series using the embedding dimension and time delay constant for each scalar time series. In addition, the neighboring points are selected using Euclidian distance. Then the new regularization constant of SVR is calculated using the weighting function whose bandwidth is optimized using the weighted distance algorithm. According to these neighboring points and the new regularization constant, the LWSVR model is set up.

Two different real-world datasets have been used to evaluate the performance of the proposed model. The proposed model has been compared with LWR, local SVR and some published papers employing the same datasets. The numerical results, achieved on the basis of different measuring errors, box plots, and Mann–Whitney U test, show the superiority of the proposed model over LWR, local SVR, and other published models. The effectiveness of the proposed model comes from weighting the SVR's regularization parameter using the LWR method where each point in the neighborhood is weighted according to its distance from the current query point. The points that are close to the current query point have larger weights than others. Moreover, by using the weighted distance algorithm, the drawback of using the weighting function's bandwidth as a fixed value has been overcome. This has led to improve the accuracy of the proposed model.

References

[1] A. D. Papalekopulos and T. C. Hesterberg, "A regression-based approach to short-term system load forecasting," *IEEE Trans. Power Syst.*, vol. 5, no. 4, pp. 1535–1547, Nov. 1990.

[2] J. W. Taylor, L. M. de Menezes, and P. E. McSharry, "A comparison of univariate methods for forecasting electricity demand up to a day ahead," *Int. J. Forecasting*, vol. 22, no. 1, pp. 1–16, 2006.

[3] J. H. Park, Y. M. Park, and K. Y. Lee, "Composite modeling for adaptive short-term load forecasting," *IEEE Trans. Power Syst.*, vol. 6, no. 2, pp. 450–457, May 1991.

[4] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. Power Syst.*, vol. 16, no. 1, pp. 44–55, Feb. 2001.

[5] M. Figueiredo, R. Ballini, S. Soares, M. Andrade, and F. Gomide, "Learning algorithms for a class of neurofuzzy network and application," *IEEE Trans. Syst., Man, Cyber. C, Appl. Rev.*, vol. 34, no. 3, pp. 293–301, Aug. 2004.

[6] C. J. Lin, C. H. Chen, and C. T. Lin, "A hybrid of cooperative particle swarm optimization and cultural algorithm for neural fuzzy networks and its prediction applications," *IEEE Trans. Syst., Man, Cyber. C, Appl. Rev.*, vol. 39, no. 1, pp. 55–68, Jan. 2009.

[7] Y. Bodyanskiy, S. Popov, and T. Rybalchenko, "Multilayer neuro-fuzzy network for short term electric load forecasting," *Lect. Notes Comput. Sci.*, vol. 5010, pp. 339–348, 2008.

[8] S. Sheng and C. Wang, "Integrating radial basis function neural network with fuzzy control for load forecasting in power system," in *Proc. IEEE Trans. Dist. Conf. Exhib. Asia and Pacific*, 2005, pp. 1–5.

[9] P. Lauret, E. Fock, R. N. Randrianarivony, and J. F. M. Ramsamy, "Bayesian neural network approach to short time load forecasting," *Energy Convers. Manage.*, vol. 49, no. 5, pp. 1156–1166, 2008.

[10] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[11] B. Scholkopf, C. Burges, and V. Vapnik, "Extracting support data for a given task," in *Proc. 1st Int. Conf. Knowl. Discovery Data Min., Menlo Park*. Menlo Park, CA: AAAI Press, 1995, pp. 252–257.

[12] B. Ribeiro, "Support vector machines for quality monitoring in a plastic injection molding process," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 401–410, Aug. 2005.

[13] J. I. Park, S. H. Baek, M. K. Jeong, and S. J. Bae, "Dual features functional support vector machines for fault detection of rechargeable batteries," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 39, no. 4, pp. 480–485, Jul. 2009.

[14] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," Royal Holloway College, Univ., London, London, U.K., NeuroCOLT Tech. Rep. NC-TR-98-030, 1998.

[15] K. W. Lau and Q. H. Wu, "Local prediction of non-linear time series using support vector regression," *Pattern Recognit.*, vol. 41, pp. 1539–1547, May 2008.

[16] C. C. Chuang and J. T. Jeng, "Selection of initial structures with support vector regression for fuzzy neural networks," *Int. J. Fuzzy Syst.*, vol. 6, no. 2, pp. 63–70, 2004.
[17] X. Li, B. Hu, and R. Du, "Predicting the parts weight in plastic injection molding using least squares support vector regression," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 6, pp. 827–833, Nov. 2008.
[18] J. A. K. Suykens, J. D. Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: Robustness and sparse approximation," *Neurocomputing*, vol. 48, no. 1–4, pp. 85–105, Oct. 2002.
[19] B. J. Chen, M. W. Chang, and C. J. Lin, "Load forecasting using support vector machines: A study on eunite competition 2001," *IEEE Trans. Power Syst.*, vol. 19, no. 4, pp. 1821–1830, Nov. 2004.
[20] M. Espinoza, J. A. K. Suykens, R. Belmans, and B. D. Moor, "Electric load forecasting using kernel based modeling for nonlinear system identification," *IEEE Control Syst. Mag., Special Issue Appl. Syst. Identification*, vol. 27, no. 5, pp. 43–57, Oct. 2007.
[21] E. E. El-Attar, J. Y. Goulermas, and Q. H. Wu, "Local prediction methods for short-term load forecasting," *Int. J. Energy Res.*, to be published.
[22] E. E. El-Attar, J. Y. Goulermas, and Q. H. Wu, "Forecasting electric daily peak load based on local prediction," in *Proc. IEEE Power Eng. Soc. Gen. Meeting (PESGM 2009)*, Calgary, Canada, Jul. 26–30, pp. 1–6.
[23] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *J. Amer. Statist. Assoc.*, vol. 83, no. 403, pp. 596–610, Sep. 1988.
[24] C. C. Atkeson, A. W. Moore, and S. Schaal, "Locally weighted learning," *Artif. Intell. Rev. (Spec. Issue Lazy Learning)*, vol. 11, pp. 11–73, Apr. 1997.
[25] World wide competition within the eunite network. (2001). [Online]. Available: http://neuron.tuke.sk/competition
[26] (2008). [Online]. Available: http://www.ee.washington.edu/class/555/elsharkawi/
index_files/Page3404.htm
[27] J. D. Farmer and J. J. Sidorowich, "Predicting chaotic time series," *Phys. Rev. Lett.*, vol. 59, pp. 845–848, 1987.
[28] M. Casdagli, "Non-linear prediction of chaotic time series," *Phys. D*, vol. 35, no. 3, pp. 335–356, 1989.
[29] J. McNames, J. A. K. Suykens, and J. Vandewalle, "Winning entry of the K.U.Leuven time series prediction competition," *Int. J. Bifurcation Chaos*, vol. 9, no. 8, pp. 1485–1500, Aug. 1999.
[30] F. E. H. Tay and L. J. Cao, "Modified support vector machines in financial time series forecasting," *Neurocomputing*, vol. 48, pp. 847–861, 2002.
[31] D. E. Lee, J. H. Song, S. O. Song, and E. S. Yoon, "Weighted support vector machine for quality estimation in the polymerization process," *Ind. Eng. Chem. Res.*, vol. 44, pp. 2101–2105, 2005.
[32] G. S. Hu, Y. Z. Zhang, and F. F. Zhu, "Short-term load forecasting based on fuzzy C-mean clustering and weighted support vector machines," in *Proc. 3rd Int. Conf. Natural Comput. (ICNC 2007)*, pp. 654–659.
[33] F. Takens, Detecting strange attractors in turbulence (Lect. Notes Comput. Sci. 898). Berlin, Germany: Springer-Verlag, 1981, pp. 366–381.
[34] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *J. Statist. Phys.*, vol. 65, pp. 579–616, May 1991.
[35] L. Cao, A. Mees, and K. Judd, "Dynamics from multivariate time series," *Phys. D*, vol. 121, pp. 75–88, Oct. 1998.
[36] O. A. Omitaomu, M. K. Jeong, A. B. Badiru, and J. W. Hines, "Online support vector regression approach for the monitoring of motor shaft misalignment and feedwater flow rate," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 5, pp. 962–970, Sep. 2007.
[37] M. P. Wand and W. R. Schucany, "Gaussian-based kernels for curve estimation and window width selection," *Can. J. Statist.*, vol. 18, pp. 197–204, 1990.
[38] H. Wang, C. Cao, and H. Leung, "An improved locally weighted regression for a converter re-vanadium prediction modeling," in *Proc. 6th World Congr. Intell. Control Autom.*, Dalian, China, Jun. 21–23, 2006, pp. 1515–1519.
[39] A. T. Lora, J. C. Riquelme, and J. L. M. Ramos, Influence of kNN-Based Load Forecasting Errors on Optimal Energy Production (Lect. Notes Comput. Sci. 2902). Berlin, Germany: Springer-Verlag, 2003, pp. 189–203.
[40] A. Sorjamaa, N. Reyhani, and A. Lendasse, *Input and Structure Selection for k-NN Approximator* (Lect. Notes Comput. Sci. 3512). Berlin, Germany: Springer-Verlag, 2005, pp. 985–992.
[41] W. C. Hong and P. F. Pai, "Predicting engine reliability by support vector machines," *Int. J. Adv. Manuf. Technol.*, vol. 28, pp. 154–161, 2006.
[42] H. Mann and D. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 1947.
[43] S. R. Abbas and M. Arif, "Electric load forecasting using support vector machines optimized by genetic algorithm," in *Proc. IEEE Multitopic Conf. (INMIC 2006)*, Dec. 23–24, pp. 395–399.
[44] X. Tao, H. Renmu, W. Peng, and X. Dongjie, "Input dimension reduction for load forecasting based on support vector machines," in *Proc. IEEE Int. Conf. Elect. Utility Deregulation, Restruct. Power Tech. (DRPT 2004)*, Hong Kong, pp. 510–514.
[45] V. H. Ferreira and A. P. A. da Silva, "Toward estimation autonomous neural network-based electric load forecasters," *IEEE Trans. Power Syst.*, vol. 22, no. 4, pp. 1554–1562, Nov. 2007.
[46] R. Ramanathan, R. Engle, C. W. J. Granger, F. V. Araghi, and C. Brace, "Short-run forecasts of electricity loads and peaks," *Int. J. Forecasting*, vol. 13, pp. 161–174, 1997.

Ehab E. Elattar (S'08) was born in Egypt in 1976. He received the B.Sc. (Hons.) and M.Sc. degrees in electrical engineering from the Department of Electrical Engineering, Menoufiya University, Shebin El-Kom, Egypt, in 1999 and 2003, respectively. Currently, he is working toward the Ph.D. degree in the Department of Electrical Engineering and Electronics, The University of Liverpool, Liverpool, U.K.

He is also with the Department of Electrical Engineering, Minufiya University. His research interests include power systems analysis and operation, artificial intelligence, support vector machine, and its applications to power systems load prediction.

John (Yannis) Goulermas (M'98) was born in Greece in 1970. He received the B.Sc. (Hons, First Class) degree in computation, and the M.Sc. and Ph.D. degrees from the University of Manchester Institute of Science and Technology (UMIST), Manchester, U.K., in 1994, 1996, and 2000, respectively.

He was in the Control Systems Centre, Department of Electrical Engineering and Electronics, UMIST, where he was engaged in the area of image processing and machine vision. He was a full-time scientific developer and later as a part-time consultant in the area of revenue and pricing modeling and optimization in industry. He was also with the University of Salford, where he was engaged in the areas of virtual reality with applications to engineering, and biomechanics and gait analysis. Since 2005, he has been in the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, U.K. His research interests include pattern recognition, mathematical optimization, and machine vision, with application areas including biomedical engineering and industrial monitoring.

Q. H. Wu (M'91–SM'97) received the M.Sc.(Eng) degree in electrical engineering from Huazhong University of Science and Technology, Wuhan, China, in 1981, and the Ph.D. degree in electrical engineering from The Queen's University of Belfast (QUB), Belfast, U.K., in 1987.

From 1981 to 1984, he was a Lecturer in electrical engineering at the University of Science and Technology. From 1987 to 1991, he was a Research Fellow and subsequently a Senior Research Fellow at QUB. In 1991, he joined the Department of Mathematical Sciences, Loughborough University, Loughborough, U.K. as a Lecturer, where he was appointed a Senior Lecturer 1995. In September, 1995, he joined The University of Liverpool, Liverpool, U.K., as the Chair of Electrical Engineering in the Department of Electrical Engineering and Electronics, where he has been the Head of the Intelligence Engineering and Automation Research Group working in the areas of systems control, computational intelligence, and electric power and energy. He has authored and coauthored more than 320 technical publications, including 135 journal papers, 20 book chapters, and 2 research monographs entitled "IP Network-Based Multi-agent Systems for Industrial Automation" and "Protective Relaying of Power Systems Using Mathematical Morphology" published by Springer. His research interests include nonlinear adaptive control, mathematical morphology, evolutionary computation, time series prediction, and power system control and operation.

Prof. Wu is a Fellow of the Institution of Engineering and Technology. He is a Chartered Engineer.