Optimal training subset in a support vector regression electric load forecasting model

JinXing Che [a,*], JianZhou Wang [b], YuJuan Tang [a]

[a] Department of Science, NanChang Institute Of Technology, NanChang, JiangXi 330099, China
[b] School of Mathematics and Statistics, LanZhou University, LanZhou, GanSu 730000, China

ARTICLE INFO

ABSTRACT

This paper presents an optimal training subset for support vector regression (SVR) under deregulated power, which has a distinct advantage over SVR based on the full training set, since it solves the problem of large sample memory complexity $O(N^2)$ and prevents over-fitting during unbalanced data regression. To compute the proposed optimal training subset, an approximation convexity optimization framework is constructed through coupling a penalty term for the size of the optimal training subset to the mean absolute percentage error (MAPE) for the full training set prediction. Furthermore, a special method for finding the approximate solution of the optimization goal function is introduced, which enables us to extract maximum information from the full training set and increases the overall prediction accuracy. The applicability and superiority of the presented algorithm are shown by the half-hourly electric load data (48 data points per day) experiments in New South Wales under three different sample sizes. Especially, the benefit of the developed methods for large data sets is demonstrated by the significantly less CPU running time.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

The load prediction is invaluable in the daily operations of a power utility. It is used for various purposes, such as price and income elasticities, energy transfer scheduling, unit commitment and load dispatch. With the emergence of load management strategies, the load prediction has played a broader role in utility operations [1]. Thus, the development of an accurate, fast, simple and robust load prediction algorithm is important to electric utilities and its customers.

As the advances in statistical learning theory, support vector regression (SVR) model has become very promising and popular due to its attractive features and profound empirical performance for small sample, nonlinearity and high dimensional data application [2–5]. Quan et al. [6] proposed a weighted least squares SVR local region algorithm for nonlinear time series. Pai and Hong [7] proposed a recurrent SVR model with genetic algorithms to forecast regional electricity load. Using a robust SVR algorithm, Zhan and Cheng [8] reported a harmonic and inter-harmonic analysis of electric power system. Hybridizing two dissimilar models, literature [9–11] pointed out that further performance improvement could be made for forecasting in the competitive market.

Based on the VC dimension theory and structural risk minimization principle, the quality and complexity of the SVR solution do not depend on the dimensionality of the input space directly. Then, the solution is optimized by solving a large-scale quadratic programming problem with linear and box constraints. The memory complexity of this problem, however, is $O(N^2)$ (N is the number of training data points). As a result, some application models of medium or large training sample size are hard to load into memory, and cannot be solved by standard SVR. Determining an optimal training subset in medium or large sample size situation is very important for generalization performance, computational efficiency, high prediction accuracy and data interpretability of SVR prediction [12]. On the other hand, redundant data not only are useless for SVR prediction but also could lead to low computational efficiency and low accuracy potentially. Thus, what redundant information should be ignored in training set has been a central topic in the areas such as statistics, pattern recognition, machine learning, and computer vision. Recently, Moustakidis and Theocharis [13] proposed an efficient filter feature selection method for achieving a satisfactory trade-off between classification accuracy and dimensionality reduction. Using an improved genetic algorithm, Yang et al. [14] and Hamdani et al. [15] presented two feature selection algorithms. Yang and Yang [16] introduced a novel condensing tree for feature selection. Liao [17] studied neural models using the smallest best feature subsets of a bladder cancer data set for classification. Past work on feature selection has emphasized

* Corresponding author. Tel.: +86 15970427685; fax: +86 15970427685.
 E-mail address: jinxingche@163.com (J. Che).

the feature extraction and classification, however, less attention has been given to the critical issue of training data set reduction and time series prediction. Only the training data points near decision boundary (namely support vectors) have impact on the final prediction model, inspired by that, we present a training data set reduction algorithm for SVR. For these reasons, optimal training subset, which represents maximum information of the full training set, is presented to supply a balanced data with relatively small training sample size for SVR. Furthermore, an approximation convexity optimization framework for computing the optimal training subset is proposed in our study, and a stopping criteria for the algorithm is established. K optimal training subset, a new algorithm put forward tentatively, is employed to obtain naturally sparse optimal training subset. Further studies on convexity will be summarized in our next study.

To show the applicability and superiority of the presented algorithm, half-hourly electric load data (48 data points per day) in New South Wales are collected. Before choosing neural networks, statistical methods and other hybrid models, the nature and intended use of electric load data should be consider carefully. The results of the comparison experiments prove that time serial forecasting and control system based on this algorithm have the following advantages: (1) Since we establish an approximation convexity optimization framework for computing the optimal training subset, the subset can extract maximum information of the full training set with minimum size. (2) Faster response capability with high precision for medium and large size training set. (3) Robust to parameter variation.

In Section 2 of this paper we present the new algorithm for forecasting, and the main steps of the method are given. Then, the possible reasons behind the proposed technique are explained. In Section 3 we introduce the research design, the data description and three performance measures. Numerical results obtained and comparisons are presented and discussed in Section 4. In Section 5 we briefly review this paper and present the future research.

2. The explicit process of the new algorithm

2.1. Nonlinear dynamics and data preprocess

The goal of prediction is to infer objects' future under their own development rules. A nonlinear prediction algorithm is employed in the present study to make predictions of electric load dynamics. In this algorithm, a single-variable time series is translated into a multi-dimensional phase space, which indicates the system state at different time. According to Takens's embedding theorem [18], a scalar time series (e.g. electric load series) y_t, where $t = 1, 2, \ldots, T$, can be reconstructed as follows

$$Y_j = [y_j, y_{j+\tau}, y_{j+2\tau}, \ldots, y_{j+(m-1)\tau}] \tag{1}$$

where $j = 1, 2, \ldots, T - (m-1)\tau$, m is the embedding dimension, τ is the time delay. Based on phase space reconstruction in a dimension m, we can construct the underlying dynamics in the form of a reflection $F : R^m \to R^m$ which makes

$$Y_{j+1} = F(Y_j) \tag{2}$$

where Y_j and Y_{j+1} are vectors of dimension m, describing the state of the system at times j (present state) and $j + 1$ (future state), respectively.

Then forecasting reflection $\bar{F} : R^m \to R^m$ can be denoted as:

$$Y_{j+1} = \bar{F}(Y_j) \tag{3}$$

That is:

$$y_{j+1} = \bar{F}(Y_j) \tag{4}$$

Here, it is nonlinear and needs to find a specific function expression, which can be obtained by utilizing ε-SVR's good ability of simulating nonlinear reflection.

Before ε-SVR can be trained with the training data, the data must be normalized over a range so as to determine the parameters' values of ε-SVR [19]. Various normalization methods are generally used for this purpose. Liu et al. [20] compared 6 different normalization methods based on the RPROP algorithm used for target recognition, they concluded that the linear normalization method and the component-whitening method have given almost the best results and are simple in concept. For the simple, we map the data linearly over a specified range $[y_{min}, y_{max}]$ in this study. Assumed that y'_{max} and y'_{min} are the maximum and minimum values of the range for the transformed variable; and y_{max} and y_{min} are the maximum and minimum values of the training data. So each value of a variable y is transformed as follows:

$$y' = P(y) = A \times y + B \tag{5}$$

Since (y_{max}, y'_{max}) and (y_{min}, y'_{min}) are two points of the above formula, so we can get the constants A and B as follows:

$$\begin{cases} A = \{y'_{max} - y'_{min}\}/\{y_{max} - y_{min}\} \\ B = y'_{max} - \{\{y'_{max} - y'_{min}\}/\{y_{max} - y_{min}\}\} \times y_{max} \end{cases} \tag{6}$$

After the employment of the proposed method, the data can revert to the un-normalized data by the following formula.

$$y = P^{-1}(y') = \frac{(y' + B)}{A} \tag{7}$$

2.2. Support vector regression

Support vector regression (SVR) [21] is a regression technique utilizing kernel functions, that is a nonlinear extension of the Generalized Portrait algorithm developed in Russia in the sixties. This subsection briefly introduces SVR, which performs the nonlinear mapping for time-series forecasting; and we refer the reader to the excellent surveys for a more thorough coverage of it [22–26].

Suppose we are given training data $(x_1, y_1), \ldots, (x_n, y_n) \subset \mathcal{W} \times \mathbb{R}$, where $\mathcal{W}$ denotes the space of the input patterns x_i (e.g. $\mathcal{W} = \mathbb{R}^n$), and y_i is the associated output values of x_i. In ε-SVR [27], our aim is to produce a function $\bar{F}(x)$ that has at most ε deviation from the actually obtained targets y_i for all training data, and simultaneously, is as "flat" as possible. That is, we do not care about errors as long as they are less than ε, but will not tolerate any deviation larger than this. Hence we arrive at the formulation stated in [27] for ε-SVR.

$$\min_{\omega, b, \xi, \xi^*} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{8}$$

$$s.t. \begin{cases} y_i - (\langle \omega, x_i \rangle + b) \le \varepsilon + \xi_i \\ (\langle \omega, x_i \rangle + b) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{9}$$

where n represents the number of samples, the constant $C > 0$ decides the trade-off between the flatness of $\bar{F}(x)$ and the amount up to which deviations larger than ε are tolerated, ξ_i denotes the upper training error, whereas ξ_i^* is the lower training error subject to ε-insensitive tube $|y_i - (\langle \omega, x_i \rangle + b)| \le \varepsilon$. This ε-insensitive loss function $|\xi_\varepsilon|$ can be described by the following equation.

$$|\xi|_\varepsilon := \begin{cases} 0, & \text{if} |\xi| < \varepsilon \\ |\xi| - \varepsilon, & \text{otherwise} \end{cases} \tag{10}$$

Instead of minimizing the observed training error, ε-SVR attempts to minimize the generalization error bound so as to achieve generalized performance, and this makes ε-SVR very robust