Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers and Chemical Engineering 28 (2004) 1489–1498
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

Soft sensing modeling based on support vector machine and Bayesian model selection

Weiwu Yan*, Huihe Shao, Xiaofan Wang

Department of Automation, Shanghai Jiao Tong University, No. 1954, Huashan Road, Shanghai 200030, China

Received 19 November 2002; received in revised form 22 September 2003; accepted 25 November 2003

Abstract

Soft sensors have been widely used in industrial process control to improve the quality of product and assure safety in production. The core of a soft sensor is to construct a soft sensing model. This paper introduces support vector machine (SVM), a new powerful machine learning method based on statistical learning theory (SLT), into soft sensor modeling and proposes a new soft sensing modeling method based on SVM. A model selection method within the Bayesian evidence framework is proposed to select an optimal model for a soft sensor based on SVM. In case study, soft sensors based on SVM are applied to the estimation of the freezing point of light diesel oil in distillation column. The estimated outputs of SVM soft sensors with the optimal model match the real values of the freezing point of light diesel oil and follow the varying trend of the freezing point of light diesel oil very well. Experiment results show that SVM provides a new and effective method for soft sensing modeling and has promising application in industrial process applications.

Keywords: Soft sensor; Modeling; Support vector machine; Distillation column

1. Introduction

It is well known that some important process variables in process control systems (e.g. product composition in a large distillation column, biomass concentration in mycelia fermentation and concentration of reaction mass in chemical reactor) are difficult or impossible to measure on-line due to the limitation of process technology or measurement techniques. These variables, which are the key indicators of process performance, are normally determined by off-line sample analyses in laboratory or on-line product quality analyzer, which is often expensive and requires frequent and high cost maintenance. Furthermore, significant delay (often several hours) will be incurred in laboratory testing such that the measured signals cannot be used as feedback signals for quality control systems. Such limitations can have a severe influence on the quality of product and safety in production. Joseph and Brosillow (1978) introduced inference control of processes to solve the problem. The basic idea of inference estimation (soft sensing) is to estimate the outputs of the primary variables by the easily measured secondary variables, which are correlated to the primary variables. In recent years, soft sensors (estimators) have been widely studied and used in industrial process control.

The core of a soft sensor is the soft sensing model of a plant, which generates a virtual measurement to replace a real sensor measurement. Generally speaking, soft sensing modeling is a problem of signal estimation, interpolation and prediction. Models based on first principles are called phenomenological models. However, the phenomenological model is often not available due to the complexity of industrial processes. As a result, empirical models are the most popular ones to develop soft sensors. The problem of empirical modeling is to find a model with the best generalization and prediction performance, given the empirical data. Currently, soft sensing modeling techniques based on empirical models include multivariate statistics, Kalman filters (KF), artificial neural networks (ANN), regression based on model, fuzzy logic and hybrid methods.

As robust methods for constructing empirical models, multivariate statistical methods such as principal component analysis (PCA) and partial least squares (PLS) have attracted wide interest (Kresta, Marlin, & MacGregor, 1994; Mejdell & Skogestad, 1991). In particular, PLS and

* Corresponding author. Tel.: +86-216-293-4831; fax: +86-216-293-2138.
E-mail address: yanwwsjtu@hotmail.com (W. Yan).

0098-1354/$ – see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2003.11.004

its variations have been applied to many practical regression problems in chemical engineering (Skagerberg, MacGrgor, & Kiprissides, 1992; Sungyong & Chonghun, 2000). However, large samples are needed in multivariate statistical methods, and models are insensitive to measurement errors.

Soft sensing models based on Kalman filter can be used when the phenomenological knowledge of the plant allows a convenient state observability (Crisafulli et al., 1996). Kalman filter is the optimal state estimator for a linear system when a model for the system together with the knowledge of certain stochastic properties of measurement and disturbance noises is available. The extended Kalman filter (EKF) is an extension of Kalman filter linear approach to nonlinear ordinary differential equations and has many successful applications (Gee & Ramirez, 1996; Gudi, Shah, & Gray, 1995). However, it should be noted that the successful application of KF and EKF depend largely on the accuracy of the process model and prior estimates of the measurement noise.

In recent years, artificial neural network has been widely used as useful tool to the nonlinear soft sensing modeling. Many soft sensing models based on different ANN architectures such as back propagation (BP) networks and radial basis function (RBF) networks have been proposed and successfully applied in industrial processes (De Assis & Filho, 2000; Glassey, Ignova, Ward, Montague, & Morris, 1997; Wang, Luo, & Shao, 1996; Yang & Chai, 1997). However, there are still no guarantees of high convergence speed, avoidance of local minima and the overfitting phenomenon, and there are no general methods to choose the number of hidden units in general neural networks.

Other methods include regression based on model (Chen, Qin & Billings, 1990; Ljung, 1987) and hybrid methods (Gomez Sanchez et al., 1999; Wang et al., 1996). Clustering has also been used in soft sensors for estimating a variable for which there is no on-line measurement in a distillation column (Espinoza, Gonzalez, Casali, & Ardiles, 1995).

In this paper, the support vector machine (SVM) (Vapnik, 1999) is employed to construct a soft sensing model. SVM is a novel powerful machine learning method based on statistical learning theory (SLT), which is a small-sample statistical theory introduced by Vapnik (1998). SVM is powerful for the problems characterized by small samples, nonlinearity, high dimension and local minima. Currently, SVM is an active field in artificial intelligent technology, and has been applied to pattern recognition, function estimation and signal processing (Suykens & Vandewalle, 1999; Vapnik, 1999). The empirical risk minimization (ERM) principle is generally employed in the classical methods such as the least-square methods, the maximum likelihood methods and traditional ANN. In SVM, the ERM is replaced by the structural risk minimization (SRM) principle, which seeks to minimize an upper bound of the generalization error rather than minimize the training error (Theodoros, Tomaso, & Massimiliano, 2002; Vapnik, 1998). Based on this principle, SVM achieves an optimum network structure by striking a right balance between the quality of the approximation of the given data and the complexity of the approximating function. Therefore, the overfitting phenomenon in general ANN can be avoided and excellent generalization performance can be obtained. Furthermore, in SVM, support vectors corresponding to the hidden units of general ANN are automatically determined after the SVM training. This implies that the difficult task of determining the network structure in general ANN can be avoided.

The paper is organized as follows. The SVM nonlinear regression algorithms including the standard SVM regression (Vapnik, 1999) and least squares SVM (LS SVM) regression algorithms (Suykens, 2001) are reviewed in Section 2. A Bayesian model selection method for the standard SVM and LS SVM regressions is introduced in Section 3. A soft sensing modeling method based on SVM is proposed in Section 4. Case study about estimation of the freezing point of light diesel oil in fractionator is given in Section 5. Finally, conclusions are given in Section 6.

2. Support vector machine regression

The basic idea of the SVM regression is to map the input data into a feature space via a nonlinear map. In the feature space, a linear decision function is constructed. The SRM principle is employed in constructing optimum decision function. Then SVM nonlinearly maps the inner product of the feature space to the original space via kernels. The SVM nonlinear regression algorithms are reviewed in this section.

Given a set of training data

$$(x_1, y_1), \ldots, (x_l, y_l) \in R^n \times R$$

The nonlinear function $\psi(\cdot)$ was employed to map original input space R^n to higher dimensional feature space R^k:$\psi(x) = (\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_l))$, where k ($k \succ\succ n$) represents the dimension of feature space. Then an optimum decision function $f(x_i) = w\varphi(x_i) + b$ is constructed in this higher dimensional feature space, where $w = (w_1, \ldots, w_k)$ is a vector of weights in this feature space. Nonlinear function estimation in the original space becomes a linear function estimation in feature space. By the SRM principle, we obtain the optimization problem:

$$\text{Minimize} \quad R = \tfrac{1}{2}||w||^2 + cR_{\text{emp}},$$

where $R_{\text{emp}} = (1/l)\sum_{i=1}^{l} L(y_i, f(x_i))$ is the error term, i.e. empirical risk in learning theory, $||w||^2$ is the regularization term, i.e. confidence interval, which controls the complexity of model (Theodoros et al., 2002), and c is a regularization parameter. $L(y_i, f(x_i))$ is the loss function which is

the loss or discrepancy between the y to a given input x and goal function $f(x)$. In the SVM regression, $L(y_i, f(x_i))$ is the ε-insensitive loss function, which generally includes the linear ε-insensitive loss function, the quadratic ε-insensitive loss function and the Huber loss function (Vapnik, 1999). Different SVM algorithms can be constructed by selecting a different ε-insensitive loss function.

2.1. Standard SVM regression algorithm

A linear ε-insensitive loss function is selected in the standard SVM regression. The optimization objective of the standard SVM regression is formulated as

$$\min J(w, \xi) = \frac{1}{2} ww + c \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{1}$$

subject to

$$y_i - w\varphi(x_i) - b \le \varepsilon + \xi_i,$$

$$w\varphi(x_i) + b - y_i \le \varepsilon + \xi_i^*,$$

$$\xi_i^*, \xi_i \ge 0, \quad i = 1, \ldots, l,$$

where ξ_i and ξ_i^* are slack variables and ε is the accuracy demanded for the approximation.

The solution to this optimization problem is given by the saddle point of the Lagrangian:

$$\begin{aligned} L(w, \xi^*, \xi, a, a^*, c, \beta, \beta^*) &= \frac{1}{2} ww + c \sum_{i=1}^{l} (\xi_i + \xi_i^*) \\ &\quad - \sum_{i=1}^{l} a_i ((w\varphi(x_i)) - y_i + b + \varepsilon + \xi_i) \\ &\quad - \sum_{i=1}^{l} a_i^* (y_i - (w\varphi(x_i)) - b + \varepsilon + \xi_i^*) \\ &\quad - \sum_{i=1}^{l} (\beta_i \xi_i + \beta_i^* \xi_i^*) \end{aligned} \tag{2}$$

(minimum with respect to elements w, b, ξ_i and ξ_i^* and maximum with respect to Lagrange multipliers $a_i \succ 0$, $a_i^* \succ 0$, $\beta_i \succ 0$ and $\beta_i^* \succ 0$, $i = 1, \ldots, l$).

From the optimality conditions

$$\frac{\partial L}{\partial w} = 0, \qquad \frac{\partial L}{\partial b} = 0, \qquad \frac{\partial L}{\partial \xi_i^*} = 0, \qquad \frac{\partial L}{\partial \xi_i} = 0, \tag{3}$$

we have

$$w = \sum_{i=1}^{l} (a_i - a_i^*) \varphi(x_i), \qquad \sum_{i=1}^{l} (a_i - a_i^*) = 0,$$

$$c - a_i - \beta_i = 0, \quad c - a_i^* - \beta_i^* = 0, \quad i = 1, \ldots, l. \tag{4}$$

Based on the Mercer's condition (Vapnik, 1999), we define kernels

$$K(x_i, x_j) = \varphi(x_i)\varphi(x_j). \tag{5}$$

By (2), (4) and (5), the optimization problem can be rewritten as

$$\begin{aligned} \max W(a, a^*) = &-\frac{1}{2} \sum_{i,j=1}^{l} (a_i - a_i^*)(a_j - a_j^*) K(x_i, x_j) \\ &+ \sum_{i=1}^{l} (a_i - a_i^*) y_i - \sum_{i=1}^{l} (a_i - a_i^*) \varepsilon \end{aligned} \tag{6}$$

subject to

$$\sum (a_i - a_i^*) = 0,$$

$$0 \le a_i \le c, \quad i = 1, \ldots, l,$$

$$0 \le a_i^* \le c, \quad i = 1, \ldots, l.$$

Finally, nonlinear function is obtained as

$$f(x) = \sum_{i=1}^{l} (a_i - a_i^*) K(x_i, x_j) + b. \tag{7}$$

2.2. Least squares SVM regression algorithm

The quadratic ε-insensitive loss function is selected in the LS SVM regression. The optimization problem of the LS SVM regression is formulated as

$$\min J(w, \xi) = \frac{1}{2} ww + c \frac{1}{2} \sum_{i=1}^{l} \xi_i^2, \tag{8}$$

subject to the equality constraints

$$y_i = w\varphi(x_i) + b + \xi_i, \quad i = 1, \ldots, l.$$

We define the Lagrangian as

$$\begin{aligned} L(w, b, \xi, a, \gamma) = &\frac{1}{2} ww + c \frac{1}{2} \sum_{i=1}^{l} \xi_i^2 \\ &- \sum_{i=1}^{l} a_i (w\varphi(x_i) + b + \xi_i - y_i), \end{aligned} \tag{9}$$

where a_i $(i = 1, \ldots, l)$ are Lagrange multipliers.

By the optimality conditions

$$\frac{\partial L}{\partial w} = 0, \qquad \frac{\partial L}{\partial b} = 0, \qquad \frac{\partial L}{\partial \xi} = 0, \qquad \frac{\partial L}{\partial a} = 0, \tag{10}$$

we have

$$w = \sum_{i=1}^{l} a_i \varphi(x_i), \qquad \sum_{i=1}^{l} a_i = 0, \qquad a_i = c\xi_i,$$

$$w\varphi(x_i) + b + \xi_i - y_i = 0. \tag{11}$$

By (5) and (11), the optimization problem can be rewritten as

$$\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & K(x_1,x_1)+\frac{1}{c} & \cdots & K(x_1,x_l) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & K(x_l,x_1) & \cdots & K(x_l,x_l)+\frac{1}{c} \end{bmatrix} \begin{bmatrix} b \\ a_1 \\ \vdots \\ a_l \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ \vdots \\ y_l \end{bmatrix}. \tag{12}$$

Finally, the nonlinear function takes the form:

$$f(x) = \sum_{i=1}^{l} a_i K(x, x_i) + b. \tag{13}$$

2.3. Kernels

Kernels $K(x, x_i)$ can be any symmetric function satisfying the Mercer's condition. Different kernels, $K(x, x_i)$, can be selected to construct different types of SVM. Typical examples include (Vapnik, 1999):

(1) polynomial kernels $K(x, x_i) = [(xx_i) + 1]^{\mathrm{d}}$;
(2) radial basic function kernels $K(x, x_i) = \exp(-((||x - x_i||^2)/2\sigma^2))$; and
(3) network kernels $K(x, x_i) = S(\upsilon(x, x_i) + c)$.

3. Bayesian model selection

In the SVM algorithm, the regularization parameter (such as c) and kernel parameters (such as σ in RBF kernels) have to be selected. These parameters play a key role in the SVM performance. The problem of choosing the values of these parameters so as to minimize the test error is called the model selection problem. These parameters are sometimes guessed by users. More disciplined approaches and methods include: validation set (Müller et al., 1997), cross-validation (Baesens et al., 2000), VC bounds (Vapnik, 1999) and Bayesian learning (Kowk, 2000; Van Gestel et al., 2001). This section aims at parameter's tuning method within the Bayesian evidence framework.

The Bayesian evidence framework, introduced by Mackay (Mackay, 1992, 1997), has been applied to the design of networks. The Bayesian evidence framework divides the inference into three distinct levels. Kowk (2000) applied the Bayesian evidence framework to the standard SVM classification algorithm. In this section, we apply the Bayesian evidence framework to the standard SVM regression algorithm and LS SVM regression algorithm in order to select optimal regularization and optimal kernel parameters. Training of the SVM regression can be statistically interpreted in level 1 inference. The optimal regularization parameter can be inferred in level 2 inference. The optimal kernel parameter selection in the SVM regression can be performed in level 3 inference.

The basic idea of model selection within the Bayesian evidence framework is to maximize the posterior probability of parameter distribution to obtain the optimal parameter. According to the Bayesian theory, the most possible parameters value and optimal model are obtained at the point where posterior of these parameters distribution are maximum. The Bayesian rule can be described as

$$\text{Posterior probability} = \frac{\text{likelihood} \times \text{prior probability}}{\text{evidence}}.$$

3.1. Level 1 inference

To be convenient, we divide optimization objective in (1) and (8) by c and then replace $1/c$ by λ. For a given value of λ, the first level of inference infers the posterior of w by

$$p(w|D, \lambda, H) = \frac{p(D|w, \lambda, H)p(w|\lambda, H)}{p(D/\lambda, H)} \propto p(D|w, \lambda, H)p(w|\lambda, H), \tag{14}$$

where D is the training data set and H represents model with parameter vector w.

Assuming training data are independently identically distributed, we obtain

$$p(D|w, \lambda, H) = \prod_{i=1}^{l} p(y_i|x_i, w, \lambda, H)p(x_i|w, \lambda, H), \tag{15}$$

where $p(x_i|w, \lambda, H)$ is a constant. We assume

$$p(y_i|x_i, w, \lambda, H) \propto \exp(-L(y_i, f(x_i)), \tag{16}$$

where $L(y_i, f(x_i))$ is the loss function and $p(w|\lambda, H)$ is the Gaussian probability distribution

$$p(w/\lambda, H) = \left(\frac{\lambda}{2\pi}\right)^{k/2} \exp\left(-\frac{\lambda}{2} w^{\mathrm{T}} w\right). \tag{17}$$

The substitution of (15), (16) and (17) in (14) provides

$$p(w|D, \lambda, H) \propto \exp\left(-\sum_{i=1}^{l} L(y_i, f(x_i))\right) \exp\left(-\frac{\lambda}{2} w^{\mathrm{T}} w\right) = \exp\left(-\frac{\lambda}{2} w^{\mathrm{T}} w - \sum_{i=1}^{l} L(y_i, f(x_i))\right). \tag{18}$$

In level 1 inference, training of SVM, i.e. (1) and (8), can be interpreted as maximizing $p(w|D, \lambda, H)$ with respect to w.

3.2. Level 2 inference

Applying the Bayes rule in the second level of inference, we obtain the posterior probability of λ

$$p(\lambda|D, H) = \frac{p(D|\lambda, H)p(\lambda|H)}{p(D|H)} \propto p(D|\lambda, H)p(\lambda|H)$$

The most possible value of λ can be determined by maximizing the posterior probability of λ as $p(\lambda|D, H)$. Assuming $p(\lambda|H)$ is flat prior distribution, we obtain

$$p(\lambda|D, H) \propto p(D|\lambda, H) \propto \int p(D|w, \lambda, H)p(w|\lambda, H)\,\mathrm{d}w \propto \left(\frac{\lambda}{2\pi}\right)^{k/2} \int \exp\left(-\frac{\lambda}{2}w^{\mathrm{T}}w - \sum_{i=1}^{l} L(y_i, f(x_i))\right)\mathrm{d}w. \tag{19}$$

Defining $E_{\mathrm{W}} = 1/2(w^{\mathrm{T}}w)$, $E_{\mathrm{D}} = \sum_{i=1}^{l} l(y_i, f(x_i))$ and $M(w) = \lambda E_{\mathrm{W}} + E_{\mathrm{D}}$, (19) can be rewritten as

$$p(\lambda|D, H) \propto \left(\frac{\lambda}{2\pi}\right)^{k/2} \int \exp(-M(w))\,\mathrm{d}w. \tag{20}$$

Taylor-expanding $M(w)$ at w_{MP}, where w_{MP} is the most possible value of parameters w, we obtain

$$M(w) = M(w_{\mathrm{MP}}) + \tfrac{1}{2}(w - w_{\mathrm{MP}})A(w - w_{\mathrm{MP}}) = \lambda E_{\mathrm{W}}^{\mathrm{MP}} + E_{\mathrm{D}}^{\mathrm{MP}} + \tfrac{1}{2}(w - w_{\mathrm{MP}})A(w - w_{\mathrm{MP}}), \tag{21}$$

where

$$A = \frac{\partial^2(E_{\mathrm{W}} + E_{\mathrm{D}})}{\partial w} = \nabla^2\left(\lambda E_{\mathrm{W}} + \sum_{i=1}^{l} l(y_i, f(x_i))\right).$$

There is no the first-order derivative term in $M(w)$ since the first-order derivatives with respect to w at most probable point w_{MP} are zero. Substituting (20) and (21) into (19), we obtain

$$p(\lambda|D, H) \propto (\lambda)^{k/2} \times \int \exp(-\lambda E_{\mathrm{W}}^{\mathrm{MP}} - E_{\mathrm{D}}^{\mathrm{MP}} - \tfrac{1}{2}(w - w_{\mathrm{MP}})A(w - w_{\mathrm{MP}}))\,\mathrm{d}w = (\lambda)^{k/2}\exp(-\lambda E_{\mathrm{W}}^{\mathrm{MP}} - E_{\mathrm{D}}^{\mathrm{MP}})(2\pi)^{k/2}\det{}^{-1/2}A, \tag{22}$$

which implies that

$$\ln p(\lambda|D, H) \propto \ln p(D|\lambda, H) = -\lambda E_{\mathrm{W}}^{\mathrm{MP}} - E_{\mathrm{D}}^{\mathrm{MP}} + \tfrac{1}{2}k\ln\lambda - \tfrac{1}{2}\ln(\det A) + \text{constant}. \tag{23}$$

Maximization of the log-posterior probability of $p(\lambda|D, H)$ with respect to λ lead to the most probable value of λ_{MP} obtained by the following equation:

$$2\lambda_{\mathrm{MP}}E_{\mathrm{W}}^{\mathrm{MP}} = \gamma, \tag{24}$$

where $\gamma = k - \lambda\,\mathrm{trace}\,A^{-1}$ is called the effective number of parameters (Kowk, 2000). Different methods are used in the standard SVM and LS SVM regressions to compute the value of γ.

In the standard SVM regression: using $L(y_i - f(x_i) - \varepsilon) = \xi_i$ and $L(f(x_i) - y_i - \varepsilon) = \xi_i^*$ in the standard SVM regression, we obtain

$$A = \nabla^2\left(\lambda E_{\mathrm{W}} + \sum_{i=1}^{l}(\xi_i + \xi_i^*)\right). \tag{25}$$

Since ξ_i and ξ_i^* have not the second-order derivative, they are replaced by

$$\xi_i = (y_i - f(x_i) - \varepsilon)s(y_i - f(x_i) - \varepsilon),$$

$$\xi_i^* = (f(x_i) - y_i - \varepsilon)s(f(x_i) - y_i - \varepsilon),$$

where $s(u) = 1/(1 + \mathrm{e}^{-u})$ (Kowk, 2000). Differentiating ξ_i and ξ_i^* with respect to w, we obtain

$$\nabla^2(\xi_i + \xi_i^*) = \frac{\partial^2(\xi_i + \xi_i^*)}{\partial w^2} = r_i\varphi(x_i)\varphi(x_i)^{\mathrm{T}}, \tag{26}$$

where

$$r_i = r(y_i - f(x_i) - \varepsilon) + r^*(f(x_i) - y_i - \varepsilon),$$

$$r(y_i - f(x_i) - \varepsilon) = (y_i - f(x_i) - \varepsilon)s''(y_i - f(x_i) - \varepsilon) + 2s(y_i - f(x_i) - \varepsilon),$$

$$r^*(f(x_i) - y_i - \varepsilon) = (f(x_i) - y_i - \varepsilon)s''(f(x_i) - y_i - \varepsilon) + 2s(f(x_i) - y_i - \varepsilon).$$

Substituting (26) into (25), we obtain

$$A = \lambda I + B, \tag{27}$$

where $B = \sum_{i=1}^{l} r_i\varphi(x_i)\varphi(x_i)^{\mathrm{T}}$.

Denoting the eigenvalues of B by ρ_m, we can obtain the ρ_m by solving $\rho_m\mu_m = \bar{K}\mu_m$ (Kowk, 2000), where μ_m is the eigenvectors and $\bar{K}$ is a $l \times l$ matrix with entries $r_i\varphi(x_i)^{\mathrm{T}}\varphi(x_j) = r_iK(x_i, x_j)$. The eigenvalues $\bar{\rho}_m$ of A are

$$\bar{\rho}_m = \begin{cases} \lambda + \rho_m & m = 1, \ldots, N \\ \lambda & m = N + 1, N + 2, \ldots, k \end{cases}, \tag{28}$$

where N ($N \le l$) denotes the number of nonzero eigenvalues of $\bar{K}$. The effective number of parameters γ of standard SVM is obtained as (Kowk, 2000)

$$\gamma = k - \lambda\,\mathrm{trace}\,A^{-1} = k - \lambda\left(\underbrace{\frac{1}{\lambda} + \cdots + \frac{1}{\lambda}}_{k-N} + \frac{1}{\lambda + \rho_1} + \cdots + \frac{1}{\lambda + \rho_N}\right) = \sum_{i=1}^{N}\frac{\rho_i}{\lambda + \rho_i}. \tag{29}$$

In the LS SVM regression: using $l(y_i - f(x_i)) = 1/2(\xi_i^2) = 1/2(y_i - w\varphi(x_i) - b)^2$ in the LS SVM regression, we obtain

$$A = \nabla^2\left(\lambda E_{\mathrm{W}} + \sum_{i=1}^{l} l(y_i - f(x_i))\right) = B + \lambda I, \tag{30}$$

where $B = \sum_{i=1}^{l}\varphi(x_i)\varphi(x_i)^{\mathrm{T}}$.

Denoting the eigenvalues of B by ρ_m, we can obtain the ρ_m by solving $\rho_m\mu_m = K\mu_m$, where μ_m is the eigenvectors

and K is a $l \times l$ matrix with entries $\varphi(x_i)^{\mathrm{T}}\varphi(x_j) = K(x_i, x_j)$. The eigenvalues $\bar{\rho}_m$ of A are

$$\bar{\rho}_m = \begin{cases} \lambda + \rho_m & m = 1, \ldots, N \\ \lambda & m = N+1, N+2, \ldots, k \end{cases}, \tag{31}$$

where N ($N \leq l$) denotes the number of nonzero eigenvalues of K. The effective number of parameters γ of LS SVM is obtained as

$$\gamma = k - \lambda \,\mathrm{trace}\, A^{-1} = \sum_{i=1}^{N} \frac{\rho_i}{\lambda + \rho_i}. \tag{32}$$

3.3. Level 3 inference

The third level of inference in the evidence framework compares the different models by examining their posterior probabilities $p(H|D) \propto p(D|H)p(H)$ and can be used to find the optimum kernel parameter. Assuming the prior probability $p(H)$ over all possible models is uniform, we obtain (Mackay, 1997)

$$p(H|D) \propto p(D|H) \propto \int p(D|\lambda, H)p(\lambda|H)\,\mathrm{d}\lambda \propto \frac{p(D|\lambda_{\mathrm{MP}}, H)}{\sqrt{\gamma}}. \tag{33}$$

Therefore

$$\ln p(H|D) = -\lambda_{\mathrm{MP}} E_{\mathrm{W}}^{\mathrm{MP}} - E_{\mathrm{D}}^{\mathrm{MP}} + \tfrac{1}{2}k \ln \lambda_{\mathrm{MP}} - \tfrac{1}{2}\ln(\det A) - \tfrac{1}{2}\ln(k - \lambda_{\mathrm{MP}}\,\mathrm{trace}\, A^{-1}) + \mathrm{constant}. \tag{34}$$

The optimum kernel parameter can be obtained by maximizing log-posterior probabilities $\ln p(H|D)$ with respect to the kernel parameter. The selection method of the kernel parameter σ in RBF kernel is given below.

To obtain the most possible value of the kernel parameter σ, we set the derivative of $\ln p(H|D)$with respect to σ to zero, i.e.

$$\frac{\partial \ln p(H|D)}{\partial \sigma} = 0. \tag{35}$$

Note that

$$\frac{\partial(\lambda_{\mathrm{MP}} E_{\mathrm{W}}^{\mathrm{MP}})}{\partial \sigma} = \lambda_{\mathrm{MP}} \sum_{i,j=1}^{l} (a_i - a_j)(a_i^* - a_j^*)\frac{\partial K_{i,j}}{\partial \sigma} = \lambda_{\mathrm{MP}} \sum_{i;j=1}^{l} (a_i - a_j)(a_i^* - a_j^*) \times \exp\left(-\frac{(x_i - x_j)^2}{2\sigma^2}\right)(x_i - x_j)^2\sigma^{-3}, \tag{36}$$

$$\frac{\partial \ln(\det A)}{\partial \sigma} = \mathrm{trace}\left(A^{-1}\left(\frac{\partial A}{\partial \sigma}\right)\right) = \mathrm{trace}\left(A^{-1}\left(\frac{\partial \bar{K}}{\partial \sigma}\right)\right), \tag{37}$$

$$\frac{\partial \ln(k - \lambda_{\mathrm{MP}}\,\mathrm{trace}\, A^{-1})}{\partial \sigma} = \frac{\lambda_{\mathrm{MP}}}{k - \lambda_{\mathrm{MP}}\,\mathrm{trace}\, A^{-1}} \frac{\partial(-\mathrm{trace}\, A^{-1})}{\partial \sigma} = \frac{\lambda_{\mathrm{MP}}}{k - \lambda_{\mathrm{MP}}\,\mathrm{trace}\, A^{-1}} \mathrm{trace}\left(A^{-2}\frac{\partial B}{\partial \sigma}\right) = \frac{\lambda_{\mathrm{MP}}}{k - \lambda_{\mathrm{MP}}\,\mathrm{trace}\, A^{-1}} \mathrm{trace}\left(A^{-2}\frac{\partial \bar{K}}{\partial \sigma}\right), \tag{38}$$

we obtain the kernel parameter by substituting (36), (37) and (38) into (35) in the standard SVM regression:

$$\sigma_{\mathrm{MP}} = \left| \left(\frac{\lambda_{\mathrm{MP}} \sum_{i;j=1}^{l} (a_i - a_j)(a_i^* - a_j^*) \times \exp(-((x_i - x_j)^2/2\sigma^2))(x_i - x_j)^2}{(\lambda_{\mathrm{MP}}/(k - \lambda_{\mathrm{MP}}\,\mathrm{trace}\, A^{-1})) \times \mathrm{trace}(A^{-2}(\partial \bar{K}/\partial \sigma)) + \mathrm{trace}(A^{-1}(\partial \bar{K}/\partial \sigma))} \right)^{1/3} \right|. \tag{39}$$

Applying the similar procedure to the LS SVM regression, we obtain the kernel parameter in the LS SVM regression:

$$\sigma_{\mathrm{MP}} = \left| \left(\frac{\lambda_{\mathrm{MP}} \sum_{i;j=1}^{l} a_i a_j \times \exp(-((x_i - x_j)^2/2\sigma^2))(x_i - x_j)^2}{\lambda_{\mathrm{MP}}/(k - \lambda_{\mathrm{MP}} \times \mathrm{trace}\, A^{-1})\mathrm{trace}(A^{-2}(\partial K/\partial \sigma)) + \mathrm{trace}(A^{-1}(\partial K/\partial \sigma))} \right)^{1/3} \right|. \tag{40}$$

Note that the absolute value of σ is employed. This is because the kernel width σ in RBF kernel should be positive, and such treatment could enhance the convergence speed of the iterative process.

4. Soft sensing modeling based on SVM

A soft sensing model based on SVM is a black-box model, which based only on input–output measurements of an industrial process. In the modeling procedure, the relationship between the input and output of the plant can be emphasized while the sophisticated inner structure is ignored. The basic structure of soft sensors based on SVM is shown in Fig. 1. Before soft sensing modeling, the secondary variables that are related with the primary variable are determined according to technologic analysis. In soft sensors, measured input variables X_{m}, manipulated variables u and measured output variables y of the plant are often the candidates of the secondary variables. The secondary variables are employed to act as the inputs of the soft sensing model, and the calculated values or long time interval sample values of the primary

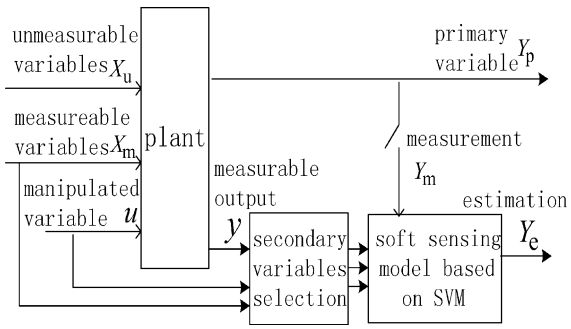

Fig. 1. Structure of soft sensor based on SVM.

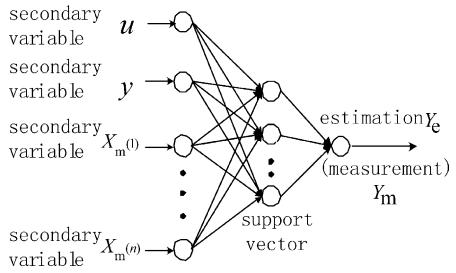

Fig. 2. Soft sensing model based on SVM.

variable Y_m are employed to act as the output of the soft sensing model. Mapping relationship of the secondary variables to the primary variable, i.e. $Y = f(u, y, X_m)$, is implemented by SVM. The soft sensing model based on SVM is shown in Fig. 2.

Procedures of soft sensing modeling based on SVM are summarized as following (Fig. 3).

- *Step 1*: the secondary variables are determined according to the theoretical analysis and experience of operators.
- *Step 2*: data are preprocessed.
- *Step 3*: kernels $K(x, x_i)$ and the SVM algorithm are determined.
- *Step 4*: given some initial guess for regularization parameter and kernel parameter, the soft sensor based on SVM are obtained through preprocessed data by using the SVM nonlinear regression algorithms.
- *Step 5*: the second level of inference determines the most possible value of the regularization parameter.
- *Step 6*: the third level of inference determines most possible value of the kernel parameter.
- *Step 7*: retrain the soft sensor based on SVM using the most possible values of the regularization and kernel parameters.
- *Step 8*: test the soft sensor based on SVM. Back to Step 4 several times and select the optimal soft sensor.
- *Step 9*: predict values of the primary variable by the optimal soft sensor based on SVM.

5. Case study: industrial distillation column

Fluid catalytic cracking unit (FCCU) is the core unit of the oil secondary operation. Its running conditions strongly affect the yield of light oil in petroleum refining. In general, FCCU consists of reactor–regenerator sub-system, fractionator sub-system, absorber–stabilizer sub-system and gas sweetening sub-system. The main aim of fractionator sub-system is to split crude oil according to a fractional distillation process. Prime products of fractionator sub-system include crude gasoline, light diesel oil and slurry. It is important that the freezing point of light diesel oil is estimated on-line in order to control the quality and yield of product.

Soft sensors based on standard SVM and LS SVM are applied to estimation of the freezing point of light diesel oil in fractionator. In this case study, the source of training and testing samples are from the process data records, which recorded and collected from the DCS systems and the corresponding daily laboratory analysis of the Shi Jia zhuang Oil Refinery Factory. Fig. 4 shows simplified flow path of fractionator.

Firstly, the secondary variables are selected according to a technology analysis. There are 32 trays in the fractionator of the Shi Jia zhuang Oil Refinery Factory. According to the analysis of the fractional distillation process, it is found that the most important contribution to the freezing point of light diesel oil is the extraction temperature of light diesel oil, which is impacted by: vapor temperature of the nineteenth tray, quantity of reflux of the first intermediate section

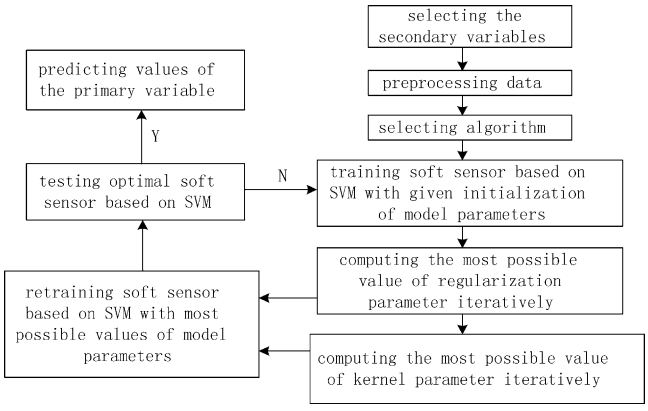

Fig. 3. Flowchart of soft sensing modeling based on SVM.

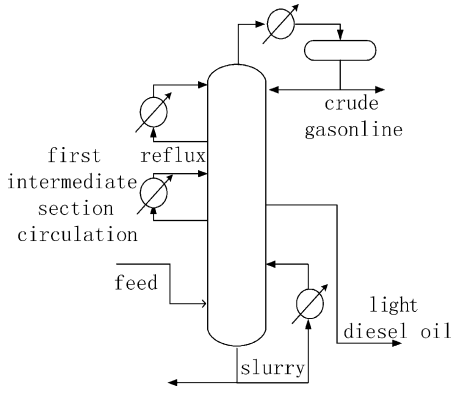

Fig. 4. Simplified flow path of fractionators.

circulation, extraction temperature and reflux temperature of first intermediate section circulation. Therefore, these five parameters are employed to act as inputs of the soft sensing model, and the freezing point of light diesel oil is employed to act as an output of the soft sensing model.

5.1. Preprocessing

Values of variables are firstly normalized to [0, 1]:

$$x = \frac{x - \min(x_i)}{\max(x_i) - \min(x_i)}, \quad i = 1, \ldots, l. \tag{41}$$

These values are further normalized to have zero mean and unit standard deviation by linear transformation (Bishop, 1995)

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_i^j, \qquad \sigma_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_i^j - \bar{x}_i)^2,$$
$$x_i^{*j} = \frac{x_i^j - \bar{x}_i}{\sigma_i}, \tag{42}$$

where x_i^* are normalized variables and j denotes the dimension.

The radial basic function is employed as kernels:

$$K(x, x_i) = \exp\left(-\frac{||x - x_i||^2}{2\sigma^2}\right), \tag{43}$$

where $||x - x_i||^2$ is calculated by formula $||x - x_i||^2 = \sum_{j=1}^{n}(x^j - x_i^j)^2$ and σ is kernels width.

5.2. Result and discussion

Fifty process data records of the secondary variables and the corresponding laboratory analysis of the freezing point of light diesel oil are selected as training samples. Using the standard SVM and LS SVM algorithms, soft sensors based on standard SVM and LS SVM are constructed through training data. Estimated outputs of the soft sensor based on standard SVM and real values of the freezing point of light diesel oil are shown in Fig. 5. Estimated outputs of the soft sensor based on LS SVM and real values of the freezing point of light diesel oil are shown in Fig. 6. Experiment results of soft sensors based on SVM are shown in Table 1.

Generalization mean squared error (GMSE) and learning mean squared error (LMSE) are general employed to evaluate the learning and generalization performance of a learning

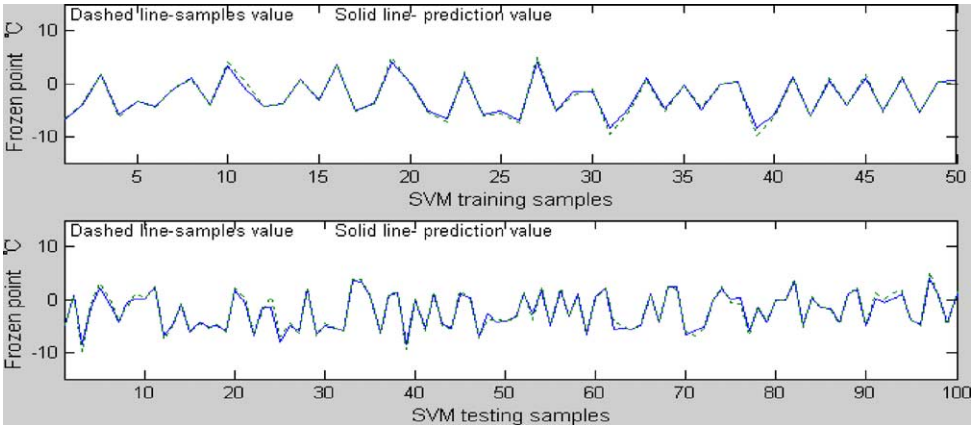

Fig. 5. Estimated outputs of the soft sensor based on standard SVM and real values of freezing point of light diesel oil ($\varepsilon = 0.1$, $\sigma = 1.8288$, $c = 0.2629$).

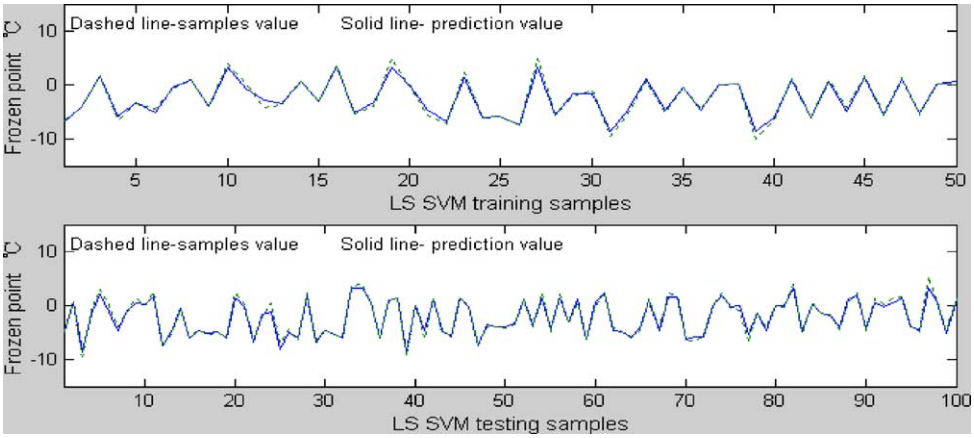

Fig. 6. Estimated outputs of the soft sensor based on LS SVM and real values of freezing point of light diesel oil ($\sigma = 0.1059$, $c = 0.9663$).

Table 1
Experiment results of soft sensors based on SVM

	The soft sensor based on standard SVM with $\varepsilon = 0.1$, $\sigma = 1.8288$ and $c = 0.2629$ (50 training samples and 100 testing samples)	The soft sensor based on LS SVM with $\sigma = 0.1059$ and $c = 0.9663$ (50 training samples and 100 testing samples)	Results of Luo and Shao (2002) (150 training samples and 50 testing samples)
LMSE	0.1204	0.1802	0.01
GMSE	0.5148	0.5476	2.4068
Number of support vectors	48	100	–

algorithm. GMSE e_{test} and LMSE e_{train} can be calculated as following:

$$e_{\text{train}} = \frac{1}{l_{\text{train}}} \sum_{(Y_{\text{m}}, Y_{\text{e}}) \in S_{\text{train}}} (Y_{\text{m}} - Y_{\text{e}})^2, \tag{44}$$

$$e_{\text{test}} = \frac{1}{l_{\text{test}}} \sum_{(Y_{\text{m}}, Y_{\text{e}}) \in S_{\text{test}}} (Y_{\text{m}} - Y_{\text{e}})^2, \tag{45}$$

where S_{train} is the training sample set, S_{test} the testing sample set, l_{test} the number of testing samples, l_{train} the number of training samples, Y_{m} the measurement value and Y_{e} is the estimation value of soft sensors based on SVM. During the model selection experimentation, it is found that a useful addition to the basic algorithm is to enforce minimum and maximum values of the model parameters. Without such bounds the algorithm occasionally get stuck in a plateau region of the model selection criterion where one or more model parameters would be very large or very small. To stop the iterative procedure, we employ the criteria that the relative error of two successive updated values is less than 0.05.

From Figs. 5 and 6, it is found that soft sensors based on SVM have good performance in estimation of the freezing point of light diesel oil. Estimated outputs of soft sensors based on SVM to the freezing point match real values of the freezing point and follow the varying trend of the freezing point very well. From Table 1, it is found that soft sensors based on SVM have good learning and generalization performance. The results of this paper are compared with that of Luo and Shao (2002), which employed ANN method. One hundred and fifty training samples are used by Luo and Shao (2002) while only 50 training samples are used in this work. The results of this work are much better than that of Luo and Shao (2002). Although the method of Luo and Shao (2002) has good learning performance, its generalization performance is not satisfactory. For small samples, excellent generalization performance is obtained in the soft sensing model based on SVM.

According to the SLT, the bound of generalization ability include two terms, i.e. risk term and confidence interval. As the structure of the model become more complex, risk term decreases and the confidence interval increases. For a given data set, a model with too complex a structure often results in poor generalization performance and the overfitting phenomena, although it may have good learning performance. SVM implements a proper balance between the quality of the approximation of the given data and the complexity of the approximating model by the SRM principle. In the proposed modeling method, the optimal model can be obtained by selecting the optimal regularization parameter in the level 2 inference and the optimal kernel parameter in the level 3 inference. The optimal model implements the well trade-off between training errors and model complexity so as to achieve well generalization performance. Therefore, for given data, the soft sensing modeling method based on SVM within the Bayesian evidence framework can find an optimum model corresponding to these data. Furthermore, support vectors in the soft sensing model base on SVM can be automatically determined after the SVM training.

Fractionator process in FCCU usually shows high nonlinearity. Experimental results and theoretical analysis show that SVM soft sensors with Bayesian model selection have the following advantages.

1. SVM is powerful for the problem with small samples. Soft sensors based on SVM are convenient for industrial process in which process parameter values are difficult to obtain.
2. SVM implements well the trade-off between the quality of the approximation of the given data and the complexity of the approximating function, the overfitting phenomena can be avoided and excellent generalization performance can be obtained in soft sensors based on SVM.
3. As a black-box model, soft sensors based on SVM are fit for industrial process in the case that a first principle model is not available.
4. Model structures of soft sensors based on SVM are automatically determined after the SVM training.
5. Bayesian model selection method provides a method to find the optimum soft sensing model based on SVM so as to achieve the excellent performance.

6. Conclusion

This paper introduces SVM into soft sensing modeling and proposes a new soft sensing modeling method based on SVM and a model selection method within the Bayesian evidence framework. SVM soft sensors within the Bayesian evidence framework are applied to estimate the freezing point of light diesel oil in distillation column. The estimated

outputs of SVM soft sensors with the optimal model to the freezing point match the real values of the freezing point and follow the varying trend of the freezing point very well. Effective results indicate that SVM modeling method within the Bayesian evidence framework provides a new tool for soft sensing modeling and has promising application in industrial process applications. Future work will focus on further application of this modeling method and investigate its stability and robustness.

Acknowledgements

This research was supported by Special Funds for Major State Basic Research of China (Project 973, Project No. G1998030415) and National High Technology R&D Program (863) Foundation of China (2001 AA413130).

References

Baesens, B., Viaene, S., Van Gestel, T., Suykens, J. A. K., Dedene, G., De Moor, B., & Vanthienen, J. (2000). An empirical assessment of kernel type performance for least squares support vector machine classification. In *Proceedings of Fourth International Conference on Knowledge-Based Intelligent Engineer system and Allied Technologies* (pp. 313–316). Brighton, UK.

Bishop, C. M. (1995). *Neural networks for pattern recognition* (pp. 298–299). Oxford: Oxford University Press.

Chen, S., Qin, S. A., & Billings, T. F. (1990). Practical identification of NARMAX models using radial basis functions. *International Journal of Control, 52*(6), 1327–1350.

Crisafulli, S., Pierce, R. D., Dumont, G. A., Ingegneri, M. S., Seldom, J. E., & Baade, C. B. (1996). Estimating sugar cane fibre rate using Kalman filtering techniques. In *Proceedings of 13th IFAC Triennial World Congress* (pp. 361–366). San Francisco, USA.

De Assis, A. J., & Filho, R. M. (2000). Soft sensors development for on-line bioreactor state estimation. *Computers and Chemical Engineering, 24*(2–7), 1099–1103.

Espinoza, P. A., Gonzalez, G. D., Casali, A., & Ardiles, C. (1995). Design of soft sensors using cluster techniques. In *Proceedings of International Mineral Processing Congress* (pp. 261–265). San Francisco, USA.

Gee, D. A., & Ramirez, W. F. (1996). On-line state estimation and parameter identification for batch fermentation. *Biotechnology Progress, 12*, 132–140.

Glassey, J., Ignova, M., Ward, A. C., Montague, G. A., & Morris, A. J. (1997). Bioprocess supervision: Neural network and knowledge based system. *Journal of Biotechnology, 52*, 201–205.

Gomez Sanchez, E., Arauzo Bravo, M. J., Cano Izquierdo, J. M., Dimitriadis, Y. A., Lopez Coronado, J., & Lopez Nieto, M. J. (1999). Control of the penicillin production using fuzzy neural networks. In *Proceedings of IEEE International Conference on SMC* (pp. 446–450). Tokyo, Japan.

Gudi, R. D., Shah, S., & Gray, M. (1995). Adaptive multirate state and parameter estimation strategies with application to a bioreactor. *America Institute of Chemical Engineering Journal, 41*(11), 2451–2464.

Joseph, B., & Brosilow, C. B. (1978). Inferential control of processes. *American Institute of Chemical Journal, 24*(3), 485–508.

Kowk, J. T. (2000). The evidence framework applied to support vector machines. *IEEE Transaction on Neural Network, 11*(5), 1162–1173.

Kresta, J. V., Marlin, T. E., & MacGregor, J. F. (1994). Development of inferential process models using PLS. *Computers and Chemical Engineering, 18*(7), 597–611.

Ljung, L. (1987). *System identification: Theory for the user. Information and system science series*. New Jersey: Prentice-Hall.

Luo, J., & Shao, H. (2002). Soft sensing modeling using neural fuzzy system based on rough set theory. In *Proceedings of 2002 American Control Conference* (pp. 543–548). Alaska, USA.

Mackay, D. J. C. (1992). Bayesian interpolation. *Neural Computing, 4*, 415–447.

Mackay, D. J. C. (1997). Probable network and plausible predictions—A review of practical Bayesian methods for supervised neural networks. *Network Computation in neural systems, 6*, 1222–1267.

Mejdell, T., & Skogestad, S. (1991). Composition estimator in a pilot-plant distillation column using multiple temperature. *Industrial and Engineering Chemistry Research, 30*, 255–2564.

Müller, K. R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., & Vapnik, V. (1997). Predicting time series with support vector machines. In *Proceedings of International Conference on Artificial Neural Networks* (pp. 999–1004). Berlin: Springer LNCS 1327.

Skagerberg, B., MacGrgor, J. F., & Kiprissides, C. (1992). Multivariate data analysis applied to low-density polyethylene reactors. *Chemometrics of Intelligent Laboratory Systems, 14*, 341–356.

Sungyong, P., & Chonghun, H. (2000). A nonlinear soft sensor based on multivariate smoothing procedure for quality estimation in distillation columns. *Computers and Chemical Engineering, 24*(2–7), 871–877.

Suykens, J. A. K. (2001). Nonlinear modeling and support vector machine. In *Proceedings of the IEEE Instrumentation and Measurement Technology Conference* (pp. 287–294). Budapest, Hungary.

Suykens, J. A. K., & Vandewalle, J. (1999). Least squares support vector machines classifiers. *Neural Network Letters, 9*(3), 293–300.

Theodoros, E., Tomaso, P., & Massimiliano, P. (2002). Regularization and statistical learning theory for data analysis. *Computational Statistics and Data Analysis, 38*(4), 421–432.

Van Gestel, T., Suykens, J. A. K., Baestaens, D. E., Lambrechts, A., Lanckriet, G., Vandaele, B., De Moor, B., & Vandewalle, J. (2001). Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Transaction on Neural Network, 12*(4), 809–821.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Vapnik, V. (1999). *The nature of statistical learning theory*. New York: Springer-Verlag.

Wang, X., Luo, R., & Shao, H. (1996). Designing a soft sensor for a distillation column with the fuzzy distributed radial basis function neural network. In *Proceedings of the 35th IEEE Conference on Decision and Control* (pp. 1714–1719). Kobe, Japan.

Yang, Y., & Chai, T. (1997). Soft sensing based on artificial neural network. In *Proceedings of the 1997 American Control Conference* (pp. 674–678). Albuquerque, USA.