Korean J. Chem. Eng., **21**(6), 1103-1107 (2004)

Weighted Support Vector Machine for Quality Estimation in Polymerization Processes

Dong Eon Lee, Sang-Oak Song and En Sup Yoon[†]

School of Chemical Engineering, Seoul National University, Seoul 151-742, Korea

(*Received 28 January 2004 • accepted 24 September 2004*)

Abstract−In this paper, a modified version of the Support Vector Machine (SVM) is proposed as an empirical model for polymerization processes modeling. Usually the exact principle models of polymerization processes are seldom known; therefore, the relations between input and output variables have to be estimated by using an empirical inference model. They can be used in process monitoring, optimization and quality control. The Support Vector Machine is a good tool for modeling polymerization process because it can handle highly nonlinear systems successfully. The proposed method is derived by modifying the risk function of the standard Support Vector Machine by using the concept of Locally Weighted Regression. Based on the smoothness concept, it can handle the correlations among many process variables and nonlinearities more effectively. Case studies show that the proposed method exhibits superior performance as compared with the standard SVR, which is itself superior to the traditional statistical learning machine in the case of high dimensional, sparse and nonlinear data.

Key words: Support Vector Machine, Smoothing, Empirical Models, Statistical Inference, Polymerization Process

INTRODUCTION

When monitoring and controlling of chemical plant processes are considered, there are some important quality variables that are difficult to measure on-line, due to the existence of certain limitations, such as cost, reliability, and long dead time. These limitations tend to cause problems in the processes themselves. These problems can be solved by using the inference model which allows those variables, which related to the qualities which we are interested in measuring such as viscosity of a polymer, to be estimated on-line by using other available on-line measurements such as temperatures and pressures. In developing the inference model, the principle or empirical model can be used, but the latter is preferred because the former is not sufficiently correct. Empirical models are usually obtained based on various modeling techniques such as multivariate statistics and artificial neural networks [Cherkassky and Mulier, 1998].

Recently, statistical learning methods have been applied to many practical problems in chemical engineering such as estimating distillation compositions [Kresta et al., 1994; Liu et al., 2000], estimating polymer quality variables [Skagerberg et al., 1992], and predicting dynamic behaviour of reaction systems [Kim and Chang, 2000].

This paper proposes a new nonlinear method which has been motivated by the Support Vector Machine (SVM) and Locally Weighted Regression (LWR).

The foundations of the SVM have been developed by Vapnik, and it has numerous attractive features and promising empirical performance compared to the traditional statistical approaches [Vapnik, 1998; Cleveland and Shawe-Taylor, 2001]. The formulation of the SVM embodies the Structural Risk Minimization (SRM) principle, which has been shown to be superior to the traditional Empirical Risk Minimization (ERM) principle, employed in conventional neural networks [Gunn et al., 1997]. It is this difference that equips SVM with a greater ability to generalize, which is one of goals in statistical learning.

LWR is motivated by the assumption that neighbouring values of the predictor variables are the best indicators of the response variable in that range of predictor values. Hence, LWR is a way of estimating a regression surface through multivariate smoothing: the response variable is smoothed as a function of the predictor variables in a moving fashion. LWR consists of developing a moving local model to a set of nearest neighbours [Cleveland and McArthur, 1988].

To increase the ability of generalization and prediction, this paper proposes the Weighted Support Vector Machine (w-SVM) for estimating the product qualities of polymerization processes with high dimensionality, nonlinearity and sparsity and these results in improved estimation accuracy for the polymerization process data.

THEORETICAL BACKGROUND

1. Support Vector Machine for Regression

Suppose there is a set of training data $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset X \times \Re$, where X denotes the space of the input patterns - for instance, $\Re^d$. SVM approximates the function with four distinct concepts: (a) Implementation of the SRM (Structural Risk Minimization) inductive principle, (b) Input samples mapped onto a very high-dimensional space using a set of nonlinear basis functions defined a priori, (c) Linear functions with constraints on complexity used to approximate the input samples in the high dimensional space, (d) The duality theory of optimization used to estimate the model parameters in a high-dimensional feature space that is computationally tractable. The use of kernel mapping enables the curse of dimensionality to be addressed. Fig. 1 conceptually illustrates the Support Vector Regression procedure.

Hence, the set of hypotheses will be a function of the type

$$f(x, w) = \sum_{i=1}^{n} w_i \phi_i(x) + b \tag{1}$$

where $\phi_i(x)$ is the point in the feature space that is nonlinearly mapped from input space x. The goal is to minimize the following risk func-

[†]To whom correspondence should be addressed.
E-mail: esyoon@pslab.snu.ac.kr

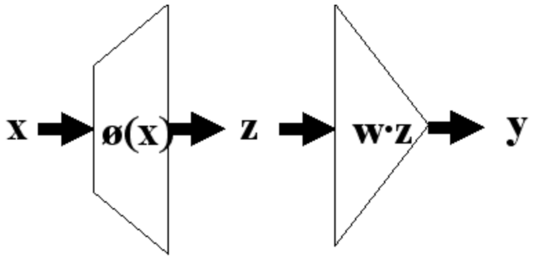

Fig. 1. Kernel mapping and regression (x: input space, z: feature space, y output space).

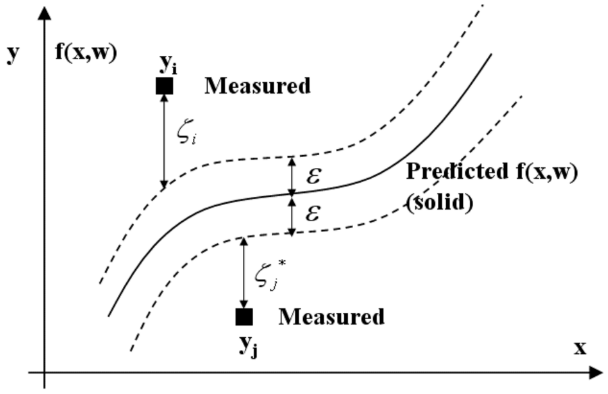

Fig. 2. The parameter used in support vector regression [Kecman, 2001].

tion:

$$\text{minimize}\frac{1}{2}\|\mathrm{w}\|^2+\mathrm{C}\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{2}$$

$$\text{subject to}\begin{cases}\mathrm{y}_i-\mathrm{w}\phi(\mathrm{x}_i)-\mathrm{b}_i\le\varepsilon+\xi_i\\ \mathrm{w}\phi(\mathrm{x}_i)+\mathrm{b}_i-\mathrm{y}_i\le\varepsilon+\xi_i^*\\ \xi_i,\xi_i^*\ge 0\end{cases} \tag{3}$$

The parameters used in Support Vector Regression are shown in Fig. 2.

The constant, C, determines the trade off between the model complexity of f and its accuracy on the training data. The formulation above corresponds to dealing with a so called ε-insensitive loss function $|\xi|_\varepsilon$ described by

$$|\xi|_\varepsilon=\begin{cases}0 & \text{if } |\xi|\le\varepsilon\\ |\xi|-\varepsilon & \text{otherwise}\end{cases} \tag{4}$$

This constrained optimization is solved by forming a primal variables Lagrangian, $\mathrm{L}_p(\mathrm{w}, \xi, \xi^*)$:

$$\begin{aligned}&\mathrm{L}_p(\mathrm{w},\mathrm{b},\xi,\xi^*,\alpha_i,\alpha_i^*,\beta_i,\beta_i^*)\\ &\quad=\frac{1}{2}\|\mathrm{w}\|^2+\mathrm{C}\sum_{i=1}^{n}(\xi_i+\xi_i^*)\\ &\quad-\sum_{i=1}^{n}\alpha_i^*(\mathrm{y}_i-\mathrm{w}\phi(\mathrm{x}_i)-\mathrm{b}-\varepsilon+\xi_i^*)\\ &\quad-\sum_{i=1}^{n}\alpha_i(\mathrm{w}\phi(\mathrm{x}_i)+\mathrm{b}-\mathrm{y}_i+\varepsilon+\xi_i)\\ &\quad-\sum_{i=1}^{n}(\beta_i^*\xi_i^*+\beta_i\xi_i)\end{aligned} \tag{5}$$

Lagrangian $\mathrm{L}_p(\mathrm{w}, \mathrm{b}, \xi, \xi^*, \alpha_i, \alpha_i^*, \beta_i, \beta_i^*)$ must be minimized with respect to the primal variables, w, b, ξ and ξ^*, and maximized with respect to non-negative Lagrange multipliers α, α^*, β and β^*. Again, this problem can be solved either in primal space or in a dual space. A solution in a dual space is chosen. It follows the Karush-Khun-Tucker (KKT) conditions for regression.

$$\partial_b\mathrm{L}_p=\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)=0 \tag{6}$$

$$\partial_w\mathrm{L}_p=\mathrm{w}-\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\mathrm{x}_i=0 \tag{7}$$

$$\partial_{\zeta_i^{(*)}}\mathrm{L}_p=\mathrm{C}-\alpha_i^{(*)}-\zeta_i^{(*)}=0 \tag{8}$$

and at the optimal solution, the product between dual variables and constraints has to vanish. This means

$$\begin{aligned}&\alpha_i(\varepsilon+\zeta_i-\mathrm{y}_i+\mathrm{w}\phi(\mathrm{x}_i)+\mathrm{b})=0\\ &\alpha_i^*(\varepsilon+\zeta_i^*-\mathrm{y}_i+\mathrm{w}\phi(\mathrm{x}_i)+\mathrm{b})=0\\ &(\mathrm{C}-\alpha_i)\zeta_i=0\\ &(\mathrm{C}-\alpha_i^*)\zeta_i^*=0\end{aligned} \tag{9}$$

Substituting (6), (7), and (8) into (5) yields the dual optimization problem.

Then the dual variables Lagrangian $\mathrm{L}_d(\alpha, \alpha^*)$ are maximized;

$$\begin{aligned}\mathrm{L}_d(\alpha,\alpha^*)&=-\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)\mathrm{y}_i\\ &\quad-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)\mathrm{K}(\mathrm{x}_i,\mathrm{x}_j)\end{aligned} \tag{10}$$

$$\text{subject to }\sum_{i=1}^{n}\alpha_i^*=\sum_{i=1}^{n}\alpha_i \tag{11}$$

$$0\le\alpha_i^*\le\mathrm{C},\quad i=1,\ldots,n$$
$$0\le\alpha_i\le\mathrm{C},\quad i=1,\ldots,n$$

where $\mathrm{K}(\mathrm{x}_i, \mathrm{x}_j)$ is the kernel function that is the inner product of point $\phi(\mathrm{x}_i)$, $\phi(\mathrm{x}_j)$ mapped into feature space. The use of kernels makes it possible to map the data implicitly into a feature space and to train a linear machine in such a space, potentially side-stepping the computational problems inherent in evaluating the feature map. With the solution of the optimization problem (11) and Eq. (9), the decision function takes the following form:

$$\mathrm{f}(\mathrm{x},\alpha_i,\alpha_i^*)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\mathrm{K}(\mathbf{x},\mathbf{x}_i)+\mathrm{b} \tag{11}$$

A list of popular kernels is shown in Table 1.

2. Locally Weighted Regression

Locally Weighted Regression is a memory-based method which is a non-parametric approach that explicitly retains the training data,

Table 1. Different types of kernel functions

Name of kernel	Expression
Polynomial degree p	$\mathrm{K}(\mathrm{x}_i, \mathrm{x}_j)=((\mathrm{x}_i\cdot\mathrm{x}_j)+1)^p$
Gaussian RBF	$\mathrm{K}(\mathrm{x}_i,\mathrm{x}_j)=\exp\left(\frac{\|\mathrm{x}_i-\mathrm{x}_j\|^2}{-2\sigma^2}\right)$
Multilayer perceptron	$\mathrm{K}(\mathrm{x}_i, \mathrm{x}_j)=\tanh((\mathrm{x}_i\mathrm{x}_j)+\mathrm{b})$

and uses it each time a prediction needs to be made. It performs a regression around a point of interest using only training data that are local to that point. This means that it is a procedure for fitting a regression surface to the data through multivariate smoothing.

To estimate g(x) of the regression surface at any value of x in the p-dimensional space of the independent variables, the ($1\le q\le n$, n: total number of observations) observations, whose x_i values are closest to x, are used. That is, a neighbourhood in the space of the independent variables is defined. Each point in the neighbourhood is weighted according to its distance from x; points close to x have a large weight, and points far from x have a small weight.

To carry out locally weighted regression, the distance function ρ in the space of the independent variables must be determined. For an independent variable, ρ, was taken to be the Euclidean distance. For multiple-regression case it is sensible to take ρ to be Euclidean distance in applications where the independent variables are measurements of position in physical space; for example, the independent variables might be the geographical location and the dependent variable might be the temperature. If the independent variables are measured on different scales, then it is typically sensible to divide each variable by an estimate of scale before applying a standard distance function.

Also, a weight function and a specification of neighbourhood size (q) must be decided on. The weight function commonly used is the following ‘tricubic’ function.

$$W(u)=(1-u^3)^3 \text{ for } \begin{cases} 0\le u\le 1 \\ 0 \quad \text{otherwise} \end{cases} \tag{12}$$

Then the weight for the observation (y_i, x_i) is

$$w_i=W(\rho(x, x_i)/d(x)) \tag{13}$$

d(x): the distance of the qth nearest x_i to x

Another weight function is the ‘Gaussian’ function.

$$W(u)=\exp(-ku^2)$$

$$\text{for } \begin{cases} 0\le u\le 1 \\ 0 \quad \text{otherwise} \end{cases} \tag{14}$$

k: smoothing parameter

Then the weight for the observation (y_i, x_i) is

$$w_i=W(\rho(x, x_i)) \tag{15}$$

Thus, $w_i(x)$, which being a function of i, has its maximum value when x_i is closest to x, decreases as the x_i increases in distance from x, and becomes 0 for the qth nearest x_i to x.

WEIGHTED SUPPORT VECTOR MACHINE

1. Formulation

As mentioned before, the constant C in Eq. (2) determines the trade off between the complexity of model, that is, f and its accuracy on the training data. When C is constant, it can be said that all training data contribute to the accuracy of the model to the same extent. However, the use of the constant C is not reasonable when the polymerization process is modeled and the output result for a certain input condition is estimated with sparse experimental data set. The model should have higher model accuracy around the training input data which are closer to the new input point for prediction. To handle this, various C, function of Euclidean distance between input data points, was used together with the concept of locally weighted regression. With this idea, the risk function can be formulated as follows:

$$\frac{1}{2}\|w\|^2+\sum_{i=1}^{n} C_i(\xi_i+\xi_i^*) \tag{12}$$

$$C_i=w_i(x_k)\times C \tag{13}$$

where x_k is input of new prediction point and $w_i(x_k)$ is the weight function obtained from (12) or (14).

With the similar procedure described in previous section replacing constant C with Eq. (13), the goal is to minimize the following dual form function with changed constraints:

$$\begin{aligned} L_d(\alpha, \alpha^*) = & -\varepsilon\sum_{i=1}^{n}(\alpha_i^*+\alpha_i)+\sum_{i=1}^{n}(\alpha_i^*-\alpha_i)y_i \\ & -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i, x_j) \end{aligned} \tag{14}$$

$$\text{subject to } \sum_{i=1}^{n}\alpha_i^*=\sum_{i=1}^{n}\alpha_i \tag{15}$$

$$0\le\alpha_i^*\le C_i, \quad i=1, \ldots, n$$
$$0\le\alpha_i\le C_i, \quad i=1, \ldots, n$$

Then the final regression function is as follows:

$$f_k(x_k, \alpha_i, \alpha_i^*)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)K(x_k, x_i)+b \tag{16}$$

where x_i(i=1, 2, …, n) is training data, x_k is new input point to predict, and f_k(g) is corresponding prediction function.

2. Conceptual Interpretation

As shown in Fig. 3, standard SVR with constant value C tries to track all training data as possible as model complexity is permitted (dashed line). It means that the amount of the prediction errors (e.g. ξ_i, ξ_i^*) is not so different. Weighted SVR gives a heavy penalty on the errors around prediction point (x mark) so it tries to reduce the errors. With this aspect, w-SVR can exhibit higher prediction performance. As the shape of the weight function is sharper, the prediction errors around new input data are decreased. But there is the possibility to overfit. The weight function is chosen though the val-

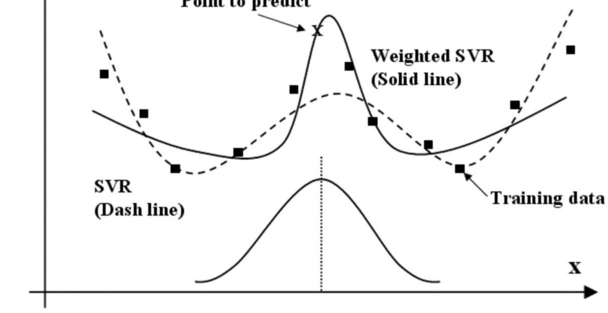

Fig. 3. Concept of weighted SVR.

idation procedure. When the location of prediction point moves, the model will be trained again but SVR will not.

CASE STUDIES

Two cases dealing with the application of w-SVM to the polymerization process are presented: one with relatively less sparse training data and the other with nonlinear and sparser data. These two cases will be used to demonstrate the performance of the proposed method.

1. Polymer Pilot Plant Data

1-1. Data Set

This data is taken from a polymer test plant [Ungar[1], 1992]. This data has previously been used for testing the robustness of nonlinear modeling methods to irregularly spaced data [DeVeaux et al., 1993]. The data set consists of 10 measurements of controlled variables in a polymer processing plant, such as temperature, feed rates and so on, and 4 measurements of the output of the plant. This data set was used just for robustness test of the proposed method so the attribute of each variable was not of concern.

1-2. Parameters

To use the weighted SVM, Euclidean distances from the input test data to each training point were calculated, and then the weighting function was calculated. Assigning a different weight penalizes each training data individually.

The 'tricubic' function was used as the weighting function, the two degree polynomial function as the kernel and the values of the parameters were chosen as follows: C=5,000 and ε=0.

1-3. Estimation and Comparison of the Results

The method was tested by using seven test data with respect to each four quality variables.

The root mean square errors (RMSE) for the test set and the relative errors (RE) are shown in Table 2. The RE is defined as follows:

$$\mathrm{RE}(\%)=\left(\frac{\mathrm{RMSE}_{SVM}-\mathrm{RMSE}_{w\text{-}SVM}}{\mathrm{RMSE}_{SVM}}\right)\times 100 \tag{16}$$

Table 2 lists the RMSE of w-SVM and RE between SVM and w-SVM. According to the relative errors, the w-SVM method reduces the prediction error of SVM by about 20%.

These results show that the w-SVM method is a robust nonlinear modeling method for irregularly spaced data and gives superior estimation performance as compared with the standard SVM method.

Table 2. Comparison of modified SVM with standard SVM using polymer test plant data

	y1	y2	y3	y4
$\mathrm{RMSE}_{w\text{-}SVM}$	0.0226	0.0188	0.0269	0.0224
RE (%)	49.9	3.09	8.19	15.6

2. PVB Experiment Data

2-1. Preliminary Process Understanding

The data is taken from the PVB (Polyvinyl Butyrate) process. The main use of PVB is in safety glass laminates, particularly in automotive, aerospace and architectural glass. Its adhesion is so strong that no glass splinters fly away when the glass laminate is broken in accidents. PVB is a polyacetal produced by the condensation of PVA (Polyvinyl Alcohol) with n-butyraldehyde in the presence of an acid catalyst [Seymour et al., 1988]. The condensation reaction produces 1,3-dioxane rings, but it is not taken to completion, leaving some unreacted hydroxyl groups which promote good adhesion to the glass substrate on lamination. Since polyvinyl alcohol is produced from the hydrolysis of polyvinyl acetate, there are a limited number of acetate groups also present. The final structure can be considered to be a random per-polymer of vinyl butyral, vinyl alcohol and vinyl acetate. Variations in chemical composition can occur depending on the reaction conditions. Therefore, the reaction conditions are normally controlled so as to impart the desired usage properties. But the principal model of the process was built with many assumptions, so predicting the quality of produced polymer with the principal model is quite different. The objective of this case study is to estimate the relationship between the control variables and the product properties using experimental data and inference models.

2-2. Data Set

This data consists of 12 measurements of controlled variables (e.g., viscosity and concentration of PVA, amount of first and second catalyst, reaction time, temperature etc.) and one measurement of the product property variable, that is, viscosity. The number of data sets is 120, but the data is sparse due to the high dimension of the input variables and the limited number of experiments. 80 out of the 120 data sets were chosen as the training set, 30 data sets were used for validation and 7 data sets were selected as the test set.

2-3. Parameters

In order to use Euclidean distance from input test data to each training point effectively, all of the input test data were normalized. The neighbourhood size (q) was set to 80 and a 'tricubic' function was used as a weighting function.

In order to make the validation error smaller, the radial basis function was used as the kernel and the values of the parameters were chosen as follows: C=1,000 and ε=0.001.

2-4. Estimation and Comparison of Results

The results of the proposed w-SVM method were compared with those of other methods, i.e., the conventional feed forward backpropagation network (FFBPN) and the standard SVM. In the case of FFBPN, the Levenberg-Marquardt backpropagation method with tangent sigmoid transfer function and three hidden layers was used. In the case of standard SVM, the same parameters and kernel function as were used with w-SVM were chosen.

Table 3. Comparison of three different methods in the prediction performance of PVB process data

	y1
RMSE_{FFBPN}	139.9667
RMSE_{SVM}	36.7946
$\mathrm{RMSE}_{w\text{-}SVM}$	23.7239
RE_{I}^{a}	83.1
RE_{II}^{b}	35.5

[a] The relative error between FFBPN and w-SVM.
[b] The relative error between SVM and w-SVM.

[1] The data and related documents are available at ftp://ftp.cis.upenn.edu/pub/ungar/chemdata

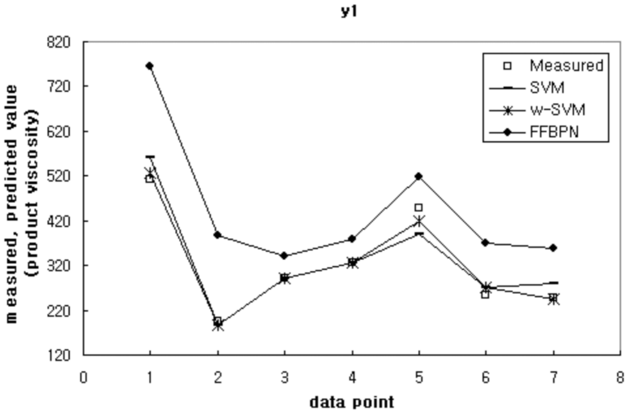

Fig. 4. Predicted and measured values in case study 2.

The RMSE for the test set and RE are shown in Table 3.

The relative error between FFBPN and w-SVM (RE_I) is 83.1% and that between standard SVM and w-SVM (RE_{II}) is 35.5%. The estimation results are shown in Fig. 4.

CONCLUSION

In this paper, we proposed a new version of the Support Vector Machine, which can be used to estimate the product properties of polymerization processes that have a highly nonlinear, high dimensional and sparse data set. To deal successfully with the nonlinearity, dimensionality and sparcity, we used the concept of Locally Weighted Regression. When it comes to minimize the error, the risk function of standard SVM attributes the same level of importance to all of the training data, but this may cause poor prediction ability when the training data set is irregularly spaced. This is the reason why we modified the standard SVM with Locally Weighted Regression. At first, the proposed method was applied to a well known data set which is frequently used for testing the robustness of nonlinear modeling methods, and then to a PolyVinyl Butyrate process data set. The result shows the improved performance of the proposed method for estimating the polymerization product properties with irregularly spaced nonlinear process data. And this model can be applied to process monitoring and optimization.

ACKNOWLEDGMENT

We acknowledge the financial aid for this research provided by the Brain Korea 21 Program, which is supported by the Ministry of Education. In addition, we would like to thank the Automation & Systems Research Institute and the Research Institute of Engineering Science of Seoul National University.

REFERENCES

Cherkassky, W. and Mulier, F., "Learning from Data," John Wiley & Sons, US (1998).

Cleveland, R. J. and McArthur, J. M., "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of American Statistical Association*, **83**, 596 (1988).

Cristianini, N. and Shawe-Taylor, J., "An Introduction to Support Vector Machines," Cambridge Univ. Press, UK (2001).

DeVeaux, R. D., Psichogios, D. C. and Ungar, L. H., "A Comparison of Two Nonparametric Estimation Schemes: MARS and Neural Networks," *Com. & Chem. Eng*., **17**(8), 819 (1993).

Gunn, S. R., Brown, M. and Bossley, K. M., "Network Performance Assessment for Neurofuzzy Data Modeling," *Intelligent Data Analysis*, **1208**, 313 (1997).

Kecman, N., "Learning and Soft Computing," MIT Press, UK (2001).

Kim, H. J. and Chang, K. S., "Hybrid Neural Network Approach in Description and Prediction of Dynamic Behaviour of Chaotic Chemical Reaction Systems," *Korean J. Chem. Eng*., **17**, 696 (2000).

Kresta, J. V., Marlin, T. E. and MacGregor, J. F., "Development of Inferential Process Models using PLS," *Com. & Chem. Eng*., **18**(7), 597 (1994).

Liu, J., Min, K., Han, C. and Chang, K. S., "Robust Nonlinear PLS Based on Neural Networks and Application to Composition Estimator for High-purity Distillation Columns," *Korean J. Chem. Eng*., **17**, 184 (2000).

Psichogios, D. C., DeVeaux, R. D. and Ungar, L. H., "Nonparametric System Identification: A Comparison of MARS and Neural Networks," Proceedings of the ACC, 1436 (1992).

Seymour, S. B. and Carraher, C. E., "Polymer Chemistry an Introduction," Marcel Dekker, New York (1988).

Skagerberg, B., MacGregor, J. F. and Kiparissides, C., "Multivariate Data Analysis Applied to Low-density Polyethylene Reactors," *Chemometrics of Intelligent Laboratory Systems*, **14**, 341 (1992).

Vapnik, V., "The Nature of Statistical Learning Theory," Springer, US (1998).