PROCESS DESIGN AND CONTROL

Weighted Support Vector Machine for Quality Estimation in the Polymerization Process

Dong Eon Lee,* Ji-Ho Song, Sang-Oak Song, and En Sup Yoon

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

In this paper, a modified version of the support vector machine (SVM) is proposed as an empirical model of polymerization processes. Polymerization processes are highly nonlinear and have a large number of input variables; hence, some qualities of their products must be estimated using an inference model rather than a principle model. The proposed method is derived by modifying the risk function of the standard SVM with the use of locally weighted regression. This method treats the correlations among the many process variables and nonlinearities, using the concept of smoothness. The case studies show that the proposed method exhibits superior performance to that of standard support vector regression, which is itself superior to the traditional statistical learning machine, in regard to treating high-dimensional, sparse, and nonlinear data.

1. Introduction

When the requirements of monitoring and controlling chemical plant processes are considered, there are always some important quality variables that are difficult to measure on-line, because of limitations such as cost, reliability, and long dead time. These limitations are also problems for the processes themselves. These problems can be solved with the use of an inference model; this approach enables the on-line estimation of those variables that are related to the qualities of interest, such as the viscosity of a polymer, from other available on-line measurements, such as the temperature and pressure. In developing an inference model, either a principle model or an empirical model can be used; however, the latter is usually preferred, because the former is insufficiently accurate. Empirical models are usually based on various modeling techniques, such as multivariate statistics and artificial neural networks.

This paper proposes a new nonlinear method that combines the support vector machine (SVM) and locally weighted regression (LWR).

The foundations of the SVM were developed by Vapnik.[1] The SVM has numerous attractive features and exhibits better empirical performance than traditional statistical approaches.

The SVM makes use of the structural risk minimization (SRM) inductive principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) inductive principle that has been used in conventional neural networks.[2] It is this difference that gives SVM a greater ability to generalize, which is one of the main goals of statistical learning.

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. Hence, 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. LWR consists of developing a moving local model to a set of nearest neighbors.

This paper proposes the weighted support vector machine (w-SVM) for estimating the product qualities of polymerization processes from data that are high-dimensional, nonlinear, and sparse, and it shows that this method provides improved accuracy in the estimation of polymerization process data.

2. Theoretical Background

2.1. Support Vector Machine for Regression. Suppose there is a set of training data $\{(x_1,y_1), ..., (x_n,y_n)\} \subset \mathrm{X} \times \mathscr{R}$, where X denotes the space of the input patterns, for instance, $\mathscr{R}^d$. The SVM approximates the underlying function with four distinct concepts:[3,4]

(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; and

(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 allows the problem of high dimensionality to be addressed. Figure 1 shows a schematic illustration of the support vector regression (SVR) procedure. Hence, the set of hypotheses will be a function of the type

$$f(x) = \sum_{i=1}^{n} w_i \phi(\mathrm{x}_i) + b \qquad (1)$$

* Corresponding author: Tel.: +82-2-873-2605. Fax: +82-2-872-1581. E-mail address: dongeon@pslab.snu.ac.kr.

10.1021/ie049908e CCC: $30.25
Published on Web 03/02/2005

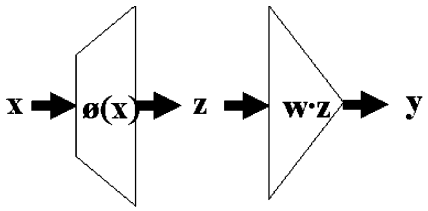

Figure 1. Kernel mapping and regression. Legend is as follows: x, input space; z, feature space; y, output space).

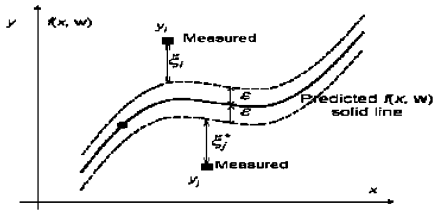

Figure 2. Parameters used in support vector regression (SVR). (From Kecman.[5])

where $\phi(x_i)$ is the point in feature space that is nonlinearly mapped from the input space x. The goal is to minimize the following risk function:

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

subject to

$$\begin{cases} y_i - w\phi(x_i) - b_i \le \epsilon + \xi_i \\ w\phi(x_i) + b_i - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{3}$$

The parameters used in SVR are shown in Figure 2.

The constant C in eq 2 determines the tradeoff between the complexity of f and its accuracy in capturing the training data. The previously described formulation is equivalent to making use of a so-called ϵ-insensitive loss function $|\xi|_\epsilon$, which is described by

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

This constrained optimization is solved by forming a primal-variables Lagrangian, $L_p(w,\xi,\xi^*)$:

$$L_p(w,b,\xi,\xi^*,\alpha_i,\alpha_i^*,\beta_i,\beta_i^*) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}\alpha_i^*(y_i - w\phi(x_i) - b - \epsilon + \xi_i^*) - \sum_{i=1}^{n}\alpha_i(w\phi(x_i) - b - y_i + \epsilon + \xi_i) - \sum_{i=1}^{n}(\beta_i^*\xi_i^* + \beta_i\xi_i) \tag{5}$$

The Lagrangian $L_p(w,b,\xi,\xi^*,\alpha_i, \alpha_i^*,\beta_i,\beta_i^*)$ must be minimized, with respect to the primal variables, w, b, ξ, and ξ^*, and also with respect to the 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 has been chosen. Applying the Karush−Khun−Tucker (KKT) conditions for regression, we maximize the dual-variables Lagrangian $L_d(\alpha,\alpha^*)$:

$$L_d(\alpha,\alpha^*) = -\epsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)y_i - \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_i^* - \alpha_i)K(x_i,x_j) \tag{6}$$

subject to

$$\sum_{i=1}^{n}\alpha_i^* = \sum_{i=1}^{n}\alpha_i \qquad (\text{for } 0 \le \alpha_i^* \le C,\ i = 1, ..., n \text{ and } 0 \le \alpha_i \le C,\ i = 1, ..., n) \tag{7}$$

where $K(x_i,x_j)$ is the kernel function that is the inner product of the points $\phi(x_i)$ and $\phi(x_j)$ mapped into feature space. The use of kernels makes it possible to map the data implicitly into a feature space and to train the linear machine in this space, with a view to sidestepping the computational problems inherent in evaluating the feature map. Finally, the decision function takes the following form:

$$f(x,\alpha_i,\alpha_i^*) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i,x) + b \tag{8}$$

A list of popular kernels is shown in Table 1.

2.2. Locally Weighted Regression. Locally weighted regression (LWR)[6] is a memory-based method that involves a nonparametric approach that explicitly retains the training data, which it uses each time a prediction must be made. Regression is performed in this method around each point of interest, using only the training data that are "local" to that point. This means that LWR is a procedure for fitting a regression surface to the data through multivariate smoothing.

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

To perform LWR, the distance function ρ in the space of the independent variables must be determined. For independent variables, ρ is usually the Euclidean distance. It is sensible to make ρ the Euclidean distance in multiple-regression applications where the independent variables are measurements of positions in physical space; for example, the independent variables might describe the geographical location and the dependent variable might be the temperature. If the independent variables are measured on different scales, then it is usually sensible to divide each variable by an estimate of scale before applying a standard distance function.

Table 1. Different Types of Kernel Functions

name of kernel	expression
polynomial degree, p	$K(x_i,x_j) = (x_ix_j + 1)^p$
Gaussian RBF[a]	$K(x_i,x_j) = \exp[\|\|x_i - x_j\|\|^2/(-2\sigma^2)]$
multilayer perceptron	$K(x_i,x_j) = \tanh((x_i,x_j) + b)$

[a] Radial basis function.

Furthermore, a weight function and a specification of neighborhood size (q) must be chosen. Cleveland and Devlin[6] introduced the M plot to determine the appropriate smoothness factor, $\boldsymbol{f}$, which is the ratio of the number of training data points used in each local regression and the number of all training data points.

The most commonly used weight function is the following "tricubic" function:

$$W(u) = \begin{cases} (1 - u^3)^3 & (\text{for } 0 \le u \le 1) \\ 0 & (\text{otherwise}) \end{cases} \quad (9)$$

The weight of observation (y_i, x_i) then is

$$w_i = W\left(\frac{\rho(x, x_i)}{d(x)}\right) \quad (10)$$

where $d(x)$ is the distance from the qth-nearest x_i to x.

Another weight function is the "Gaussian" function:

$$W(u) = \begin{cases} \exp(-ku^2) & (\text{for } 0 \le u \le 1) \\ 0 & (\text{otherwise}) \end{cases} \quad (11)$$

where k is the smoothing parameter.

The weight of the observation (y_i, x_i) then is

$$w_i = W(\rho(x,x_i)) \quad (12)$$

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

2.3. Weighted Support Vector Machine. 2.3.1. Formulation. As mentioned previously, the constant C in eq 2 determines the tradeoff between the complexity of the model (that is, $\boldsymbol{f}$) and its accuracy in capturing the training data. When C is constant, all training data contribute to the accuracy of the model to the same extent. However, the use of a constant value for C is not reasonable when a polymerization process is modeled and the output results for certain input conditions are estimated using a sparse experimental data set. The model should have higher accuracy for the training input data that are closer to the new input point for prediction. To achieve this aim, C is treated as a function of the Euclidean distance between input data points, and the concept of LWR is used. For this approach, the risk function can be formulated as follows:

$$\frac{1}{2}||w||^2 + \sum_{i=1}^{n} C_i(\xi_i + \xi_i^*) \quad (13)$$

$$C_i = w_i(x_0) \times C \quad (14)$$

where $w_i(x_0)$ is the weight function obtained from eq 10 or 12.

With a similar procedure to that described in the previous section, i.e., replacing constant C with eq 14, the goal is to minimize the following dual-form function with changed constraints:

$$L_d(\alpha,\alpha^*) = -\epsilon\sum_{i=1}^{n}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)y_i - \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i,x_j) \quad (15)$$

subject to

$$\sum_{i=1}^{n}\alpha_i^* = \sum_{i=1}^{n}\alpha_i \qquad (\text{for } 0 \le \alpha_i^* \le C_i,\ i = 1, ..., n \text{ and } 0 \le \alpha_i \le C_i,\ i = 1, ..., n) \quad (16)$$

The final regression function then is as follows:

$$f_k(x_k,\alpha_i,\alpha_i^*) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_k,x_i) + b \quad (17)$$

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

2.3.2. Conceptual Interpretation. As shown in Figure 3, the standard SVR with constant C tries to track all training data with a specific model complexity (shown by the dashed line). This means that the size of the prediction errors (e.g., ζ_i, ζ_i^*) does not vary greatly. Weighted SVR applies a heavy penalty to the errors near the prediction point (marked as "x" in Figure 3) in an attempt to reduce such errors. With this approach, w-SVR is expected to exhibit higher prediction performance, because as the shape of the weight function becomes sharper, the prediction errors around new input data are expected to decrease. However, the possibility of overfit remains. The weight function is chosen through the validation procedure. In contrast to SVR, when the location of the prediction point moves, the model is retrained.

3. Case Studies

Two examples of the application of w-SVM to polymerization processes are now presented: one with a relatively less-sparse training data set and the other with a nonlinear and sparser training set. These two cases are used to test the proposed method.

3.1. Polymer Test Plant. 3.1.1. Data Set. These data are taken from a polymer test plant.[7] These data have previously been used to test the robustness of nonlinear modeling methods with irregularly spaced data (DeVeaux et al., 1993).[8] The data set consists of

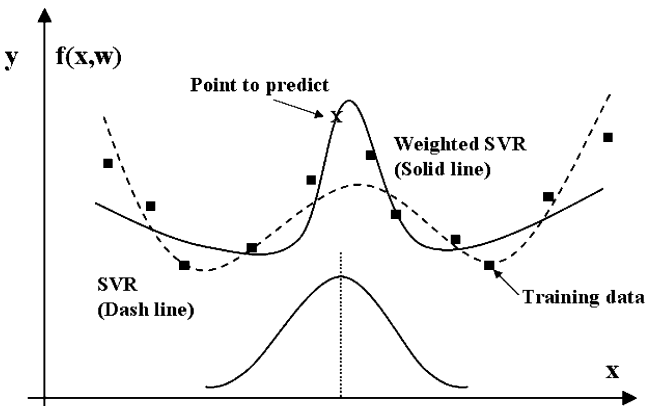

Figure 3. Weighted SVR.

Table 2. Comparison of Weighted Support Vector Machine (SVM) with Standard SVM for the Polymer Test Plant Data

	y_1	y_2	y_3	y_4
$RMSE_{w-SVM}$	0.0226	0.0188	0.0269	0.0224
RE (%)	49.9	3.09	8.19	15.6

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 to test robustness of the proposed method; therefore, the attribute of each variable was a matter of little concern.

3.1.2. Parameters. To use the weighted SVM, the Euclidean distances from the input test data point to each training data point were calculated and then the weighting function was calculated. Assigning a different weight penalizes each training data point individually. The "tricubic" function was used as the weighting function, a two-degree polynomial function was used as the kernel, and the values of the parameters were chosen as follows: $C = 5000$ and $\epsilon = 0$. Because there is no structural method for choosing the optimal parameters of SVM, the values of the parameters that produced the best results for the validation data set were used.

3.1.3. Estimation and Comparison of the Results. The method was tested using seven test data points, with respect to each of the four quality variables.

The root-mean-square error (RMSE) values for the test set and the relative error (RE) values are shown in Table 1. The relative error is defined as follows:

$$\mathrm{RE}\ (\%) = \left(\frac{\mathrm{RMSE}_{\mathrm{SVM}} - \mathrm{RMSE}_{\mathrm{w-SVM}}}{\mathrm{RMSE}_{\mathrm{SVM}}}\right) \times 100 \quad (18)$$

Table 2 lists the RMSE of w-SVM and the RE between SVM and w-SVM. According to the relative errors, the w-SVM method reduces the prediction error of SVM by ~20%. The w-SVM gives better results in the prediction of output y_1 and y_4. On the other hand, there is no noticeable difference between standard SVM and w-SVM in the prediction of output y_2 and y_3. The output variables, y_1 and y_4, have higher nonlinearity, with respect to the input variables, and exhibit sharp changes within certain input regions.

These results show that the w-SVM method is a robust nonlinear modeling method for treating irregularly spaced data and gives superior estimation performance to that of the standard SVM method when (i) the output value changes sharply with the input region, (ii) the region of the input data set is wide, and (iii) the data set is sparse.

3.2. PVB Experimental Data. 3.2.1. The PVB Process. The data set was obtained from a polyvinyl butyrate (PVB) 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 poly(vinyl alcohol) (PVA) with *n*-butyraldehyde in the presence of an acid catalyst.[9] The condensation reaction produces 1,3-dioxane rings; however, it is not taken to completion, leaving some unreacted hydroxyl groups that promote good adhesion to the glass substrate on lamination. Since poly(vinyl 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, to impart the desired properties. However, the principle model of the process was built using many assumptions; hence, predicting the quality of the produced polymer with the principle model is quite difficult. The objective of this case study is to estimate the relationship between the controlled variables and the product properties from the experimental data.

Table 3. Comparison of Three Different Methods in the Prediction of Polyvinyl Butyrate (PVB) Process Data

parameter	value
root mean square error (RMSE) of y	
$RMSE_{FFBPN}$	109.4431
$RMSE_{SVM}$	34.9047
$RMSE_{w-SVM}$	23.9254
relative error (RE) of y	
between FFBPN and w-SVM, RE_I	78.12%
between SVM and w-SVM, RE_{II}	31.5%

3.2.2. Data Set. This data set consists of 12 measurements of controlled variables (e.g., viscosity and concentration of PVA, quantities of the first and second catalysts, reaction time, temperature, etc.) and 1 measurement of the product property variable (that is, the viscosity). The number of data points is 120; however, the data set is sparse, because of the large number of input variables and the limited number of experiments. From the 120 data points, 80 were chosen as the training set, 20 were used for validation, and 20 were selected as the test set.

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

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: $C = 1000$ and $\epsilon = 0.001$.

3.2.4. 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 back-propagation network (FFBPN) and the standard SVM. In the case of FFBPN, the Levenberg−Marquardt back-propagation 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 those used for the w-SVM application were chosen.

The RMSE for the test set and RE are shown in Table 3. The relative error between FFBPN and w-SVM (RE_I) is 78.12% and that between standard SVM and w-SVM (RE_{II}) is 31.5%.

The estimation results are shown in Figure 4. The data set used in this case study has features for which w-SVM is expected to exhibit better prediction ability. The range of the input data set is wide and the viscosity value changes sharply as some input variables (the quantity of the second catalyst and the reaction time) change. The w-SVM provides increased prediction ability for this data set: the same parameters were used in the SVM and w-SVM calculations, with improved

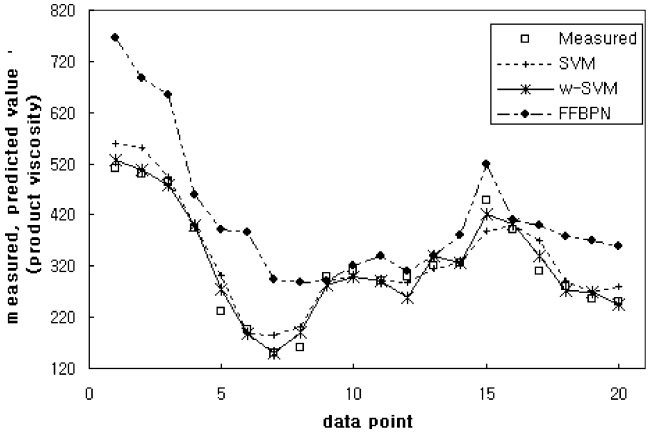

Figure 4. Predicted and measured values for case study 2.

results. These results show the effectiveness of using a weighted risk function.

4. Conclusion

In this paper, we have proposed a new version of the support vector machine (SVM), 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 sparsity of such data sets, we have applied the concept of locally weighted regression (LWR). The risk function of standard SVM attributes the same level of importance to all of the training data; however, this may result in poor prediction ability when the training data set is irregularly spaced. Thus, we have modified the standard SVM with LWR. The proposed method was applied to a well-known data set that is frequently used for testing the robustness of nonlinear modeling methods, and then to a polyvinyl butyral process data set. The results demonstrate the improved performance of the proposed method for estimating polymerization product properties with irregularly spaced nonlinear process data. This model could also 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 and Systems Research Institute and the Research Institute of Engineering Science of Seoul National University.

Literature Cited

(1) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1998.

(2) Gunn, S. R.; Brown, M.; Bossley, K. M. Network Performance Assessment for Neurofuzzy Data Modeling. In *Advances in Intelligent Data Analysis*; Lecture Notes in Computer Science, Volume 1280; Liu, X., Cohen, P., Berthold, M, Eds.; Springer: Berlin, New York, 1997; pp 313−323. (ISBN No. 3-540-63346-4.)

(3) Cherkassky, W.; Mulier, F. *Learning from Data*; Wiley: New York, 1998.

(4) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, U.K., 2001.

(5) Kecman, N. *Learning and Soft Computing*; MIT Press: London, 2001.

(6) Cleveland, R. J.; Devlin, S. J. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J. Am. Stat. Assoc.* **1988**, *83*, 596−610.

(7) Ungar, L. H. ftp://ftp.cis.upenn.edu/pub/ungar/chemdata.

(8) DeVeaux, R. D.; Psichogios, D. C.; Ungar, L. H. A Comparison of Two Nonparametric Estimation Schemes: MARS and Neural Networks. *Comput. Chem. Eng.* **1993**, *17* (8), 819−837.

(9) Seymour, S. B.; Carraher, C. E. *Polymer Chemistry—An Introduction*; Marcel Dekker: New York, 1988.

Received for review February 2, 2004
Revised manuscript received December 3, 2004
Accepted December 6, 2004

IE049908E