Anal. Chem. **2004,** *76,* 3099−3105

Multivariate Calibration with Least-Squares Support Vector Machines

Uwe Thissen, Bülent Üstün, Willem J. Melssen, and Lutgarde M. C. Buydens*

Laboratory of Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

This paper proposes the use of least-squares support vector machines (LS-SVMs) as a relatively new nonlinear multivariate calibration method, capable of dealing with ill-posed problems. LS-SVMs are an extension of "traditional" SVMs that have been introduced recently in the field of chemistry and chemometrics. The advantages of SVM-based methods over many other methods are that these lead to global models that are often unique, and nonlinear regression can be performed easily as an extension to linear regression. An additional advantage of LS-SVM (compared to SVM) is that model calculation and optimization can be performed relatively fast. As a test case to study the use of LS-SVM, the well-known and important chemical problem is considered in which spectra are affected by nonlinear interferences. As one specific example, a commonly used case is studied in which near-infrared spectra are affected by temperature-induced spectral variation. Using this test case, model optimization, pruning, and model interpretation of the LS-SVM have been demonstrated. Furthermore, excellent performance of the LS-SVM, compared to other approaches, has been presented on the specific example. Therefore, it can be concluded that LS-SVMs can be seen as very promising techniques to solve ill-posed problems. Furthermore, these have been shown to lead to robust models in cases of spectral variations due to nonlinear interferences.

The importance of multivariate calibration (MVC) methods in the field of analytical chemistry is indisputable. MVC is often used in a wide variety of industrial applications (e.g., in food, petrol, or pharmaceutical industries) to relate easily measured spectra to specific parameters of interest. This is especially useful if it is difficult to measure the parameters of interest in a direct way. For example, the characteristics (and the quality) of industrial products can often only be determined in a laborious and expensive way; therefore, these have to be measured off-line. Using MVC, these parameters can also be derived from indirect measurements, such as spectra, only much faster. Very often, Raman or near-infrared (NIR) spectra are used. Hence, the use of MVC is a very suitable approach to be used on-line for product quality estimation.

A typical characteristic of spectral data is that the variables (wavelengths or wavenumbers) are often correlated. Furthermore, usually many variables are recorded that exceed the number of spectra (i.e., the number of measurements). Therefore, performing regression with few measurements as compared to the number of variables leads to a so-called ill-posed problem. As a result, standard linear regression breaks down, implying that no solution can be obtained. Suitable candidates for regression methods on spectra should be able to deal with these problems.

The most commonly used MVC technique for laboratory and industrial purposes is partial least-squares (PLS). PLS solves the ill-posed problem by performing regression on a new basis, which is a linear combination of the original variables. Much research has been conducted on PLS to study important calibration issues such as feature selection, model transferability, or its robustness to known and unknown external interferences.[1−8]

The advantages of PLS are that it is easy to use; it is fast; its basis allows some interpretation of underlying relationships present in the data; and to some extent, it can model weak nonlinearities. However, it has been shown that PLS is not necessarily the best-performing approach,[9] especially if nonlinear calibration has to be performed.[10,11]

Recently, support vector machines (SVMs) have been introduced as promising alternatives to the existing linear and nonlinear MVC approaches.[12−15] Originally, SVMs have been developed for binary classification, but their principles can be extended for

* To whom correspondence should be addressed. Telephone: +31 24 36 53192. Fax: +31 24 36 52653. E-mail: lbuydens@sci.kun.nl.

(1) Wülfert, F.; Kok, W. Th.; Smilde, A. K. *Anal. Chem.* **1998**, *70*, 1761−1767.
(2) Swierenga, H. Robust Multivariate Calibration Models in Vibrational Spectroscopic Applications. Ph.D. Thesis, University of Nijmegen, Nijmegen, 2000.
(3) Witjes, H.; Van den Brink, M.; Melssen, W. J.; Buydens, L. M. C. *Chemom. Intell. Lab. Syst.* **2000**, *52*, 105−116.
(4) Hageman, J. A.; Streppel, M.; Wehrens, R.; Buydens, L. M. C. *J. Chemom.* **2003**, *17*, 427−437.
(5) Estienne, F.; Massart, D. L. *Anal. Chim. Acta* **2001**, *450*, 123−129.
(6) Gusnanto, A.; Pawitan, Y.; Huang, J.; Lane, B. *J. Chemom.* **2003**, *17*, 174−185.
(7) Felipe-Sotelo, M.; Andrade, J. M.; Carlosena, A.; Prada, D. *Anal. Chem.* **2003**, *75*, 5254−5261.
(8) Pérez Pavón, J. L.; del Nogal Sánchez, M.; García Pinto, C.; Fernández Laespada, M. E.; Moreno Cordero, B. *Anal. Chem.* **2003**, *75*, 6361−6367.
(9) Wentzell, P. D.; Vega Montoto, L. *Chemom. Intell. Lab. Syst.* **2003**, *65*, 257−279.
(10) Centner, V.; Verdú-Andrés, J.; Walczak, B.; Jouan-Rimbaud, D.; Despagne, F.; Pasti, L.; Poppi, R.; Massart, D. L.; De Noord, O. E. *Appl. Spectrosc.* **2000**, *54*, 608−623.
(11) Despagne, F.; Massart, D. L.; Chabot, P. *Anal. Chem.* **2000**, *72*, 1657−1665.
(12) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.
(13) Vapnik, V. *Statistical Learning Theory*; John Wiley & Sons: New York, 1998.
(14) Smola, A. J.; Schölkopf, B. *A Tutorial on Support Vector Regression*; NeuroCOLT Technical Report NC-TR-98-030; Royal Holloway College, University of London: London, 1998.
(15) Schölkopf, B.; Smola, A. J. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.

10.1021/ac035522m CCC: $27.50
Published on Web 05/05/2004

regression purposes. Yet, in the field of analytical chemistry or chemometrics only a few applications of SVM regression have been reported.[16−21]

SVMs have the advantage that these can deal with ill-posed problems and lead to global models that are often unique. Furthermore, due to their specific formulation, sparse solutions can be found, and both linear and nonlinear regression can be performed. However, finding the final SVM model can be computationally very difficult because it requires the solution of a set of nonlinear equations (quadratic programming). As a simplification of this approach, Suykens and co-workers[22,23] have proposed the use of least-squares SVM (LS-SVM). LS-SVM has been proposed as a class of kernel machines related to many other well-known techniques (e.g., kernel Fisher discriminant analysis, principal component analysis, canonical correlation analysis, PLS, or recurrent neural networks). It is also closely related to Gaussian processes and regularization networks but uses an optimization approach as in SVMs. Therefore, LS-SVM encompasses similar advantages as SVM, but its additional advantage is that it requires solving a set of only linear equations (linear programming), which is much easier and computationally very simple.

The main goal of this paper is to demonstrate the use of LS-SVM as a relatively new multivariate calibration technique. As an example case, a well-known important analytical chemical problem is used. The problem considered deals with the use of experimental data for multivariate calibration that are affected by unavoidable (nonlinear) interferences. To obtain reliable calibration models, it is important to obtain models that are robust against those interferences. One specific example of this problem has been introduced by Wülfert et al.[1] in which NIR spectra of a ternary mixture are affected nonlinearly by temperature-induced spectral variations. In the literature, various approaches have been described to solve this specific problem, but promising results have only been obtained using nonlinear regression approaches.[17,24−28] Because the reported results are still not completely satisfying, the second goal of this paper is to contribute to the solution of this problem using LS-SVM. The excellent performance of LS-SVM has been demonstrated and compared with the published results.

THEORY

The theory of LS-SVMs and its predecessor have been described clearly in the following references: Suykens et al.[23] and Schölkopf and Smola,[15] respectively. For this reason, this section only shows the important elements of performing multivariate calibration with LS-SVM using these references. First, in general its relation is discussed with standard statistical methods from the point of view of solving ill-posed problems. These kinds of problems occur if fewer measurements are taken (i.e., less objects) than the number of variables or if the variables measured are (strongly) correlated. This is typically the case when spectra are measured. Next, the approach for linear and nonlinear regression is explained, followed by a discussion of the important characteristics and advantages of LS-SVMs.

In general, there are two ways to solve ill-posed problems.[29,30] The first way is to perform regression on a basis with a lower dimension than the original one. Well-known examples are PLS or principal component regression (PCR) that use PLS factors or PCs to define the new basis, respectively. The second way to solve ill-posed problems is to shrink the regression coefficients by imposing a penalty on their values. A well-known regression method making use of this approach is ridge regression[31] (RR).

In principle, LS-SVM always fits a linear relation ($y = wx + b$) between the regressors (x) and the dependent variable (y). Similar to RR, the best relation is the one that minimizes the cost function (Q) containing a penalized regression error term:

$$Q_{\text{LS-SVM}}: \frac{1}{2}||w||^2 + \gamma\sum_{i=1}^{n} e_i^2 \text{ subject to } y_i - w^T x_i - b = e_i \quad (1)$$

The first part of this cost function is a so-called L_2 norm on the regression weights. Using this norm, weight values are penalized quadratically, and it aims at coefficients that are as small as possible. The second term takes into account the regression error (e_i) for all of the n training objects (the standard least-squares error approach). The relative weight of this part as compared to the first part is indicated by the parameter γ, which has to be optimized by the user. The third part gives the definition of the regression error to be the difference between the true and predicted values, and this can be seen as a constraint. For comparison, note that the traditional SVM approach defines the regression error differently by neglecting all regression errors smaller than $\pm\epsilon$ (the ϵ-insensitive loss function[12]). It is this difference in error definitions that makes the LS-SVM optimization problem computationally much easier than the original SVM problem. Furthermore, the value of parameter ϵ does not have to be optimized for LS-SVM, which is the case for SVMs.

The crucial difference between RR and LS-SVM depends on the approach followed to solve the optimization of the cost function. RR solves this problem by simply setting the first derivative to zero. The LS-SVM approach considers this problem to be a constrained optimization problem and uses a Lagrangian function to solve it:

(16) Song, M.; Breneman, C. M.; Bi, J.; Sukumar, N.; Bennet, K. P.; Cramer, S.; Tugcu, N. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1347−1357.
(17) Thissen, U.; Pepers, M.; Üstün, B.; Melssen, W. J.; Buydens, L. M. C. *Chemom. Intell. Lab. Syst.*, in press.
(18) Thissen, U.; van Brakel, R.; de Weijer, A. P.; Melssen, W. J.; Buydens, L. M. C. *Chemom. Intell. Lab. Syst.* **2003**, *69*, 35−49.
(19) Belousov, A. I.; Verzakov, S. A.; von Frese, J. *Chemom. Intell. Lab. Syst.* **2002**, *64*, 15−25.
(20) Belousov, A. I.; Verzakov, S. A.; von Frese, J. *J. Chemom.* **2002**, *16*, 482−489.
(21) Lukas, L.; Devos, A.; Suykens, J. A. K.; Vanhamme, L.; van Huffel, S.; Tate, A. R.; Majós, C.; Arús, C. *ESANN 2002 Proceedings Bruges*, 2002; pp 131−136.
(22) Suykens, J. A. K.; Vandewalle, J. *Neural Process. Lett.* **1999**, *9*, 293−300.
(23) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: Singapore, 2002.
(24) Wülfert, F.; Kok, W. Th.; De Noord, O. E.; Smilde, A. K. *Chemom. Intell. Lab. Syst.* **2000**, *51*, 189−200.
(25) Wülfert, F.; Kok, W. Th.; De Noord, O. E.; Smilde, A. K. *Anal. Chem.* **2000**, *72*, 1639−1644.
(26) Swierenga, H.; Wülfert, F.; De Noord, O. E.; De Weijer, A. P.; Smilde, A. K.; Buydens, L. M. C. *Anal. Chim. Acta* **2000**, *411*, 121−135.
(27) Marx, B. D.; Eilers, P. H. C. *J. Chemom.* **2000**, *16*, 129−140.
(28) Eilers, P. H. C.; Marx, B. D. *Chemom. Intell. Lab. Syst.* **2003**, *66*, 159−174.
(29) Frank, I. E.; Friedman, J. H. *Technometrics* **1993**, *35*, 109−135.
(30) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer-Verlag: New York, 2001.
(31) Hoerl, A. E.; Kennard, R. W. *Technometrics* **1970**, *12*, 55−67.

$$L(w, b, e, \alpha) = \frac{1}{2}||w||^2 + \gamma\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n}\alpha_i\{w^T x_i + b + e_i - y_i\} \quad (2)$$

In this Lagrangian (L), the first two parts are the cost function as defined earlier, but the Lagrangian is extended with the constraint multiplied by so-called Lagrange multipliers (α_i). Each Lagrange multiplier corresponds to a certain training point. To obtain the final LS-SVM solution, the partial first derivatives of this Lagrangian function are taken and are set to zero. For further details about this approach, the reader is referred to the literature.[23] However, an important subresult of this approach is that the weight coefficients (w) can be written as an expansion of the Lagrange multipliers with the corresponding training objects (x_i):

$$w = \sum_{i=1}^{n}\alpha_i x_i \text{ with } \alpha_i = 2\gamma e_i \quad (3)$$

Using the Lagrangian, the approach comes down to finding values for the Lagrange multipliers that solve the problem rather than finding the weight (w) as in RR. So, when filling in this expression into the original regression line ($y = wx + b$), the following result is obtained:

$$y = \sum_{i=1}^{n}\alpha_i x_i^T x + b = \sum_{i=1}^{n}\alpha_i\langle x_i, x\rangle + b \quad (4)$$

where the inner product of x_i and x is indicated by $\langle x_i, x\rangle$. From Suykens et al.,[23] it can be seen that the Lagrange multipliers can be defined as

$$\alpha_i = (x_i^T x_i + (2\gamma)^{-1})^{-1}(y_i - b) \quad (5)$$

Finding these Lagrange multipliers is very simple as opposed to the SVM approach in which a more difficult relation has to be solved to obtain these values. As can be seen from eqs 3 and 5, usually all Lagrange multipliers are nonzero, which means that all training objects contribute to the solution (these are all support vectors). Furthermore, training objects that are located far away from the regression line (relatively high prediction errors) highly influence the location of this line. For this reason, the corresponding Lagrange multipliers are also relatively high (proportional to their prediction error). As discussed before, SVMs neglect all regression errors of the training objects that are smaller than ϵ. As a result, their corresponding Lagrange multipliers are zero, which means that a sparse solution can be obtained: the final result only depends on a fraction of the training objects.

The advantage of solving the optimization problem in terms of the Lagrange multipliers is that the final model can be written as a weighted linear combination of the inner product between the training points and a new test object (x). The entry of the data in inner products is very important because of two reasons. The first one is that the dimension of the objects (i.e., the number of variables) does not appear in the problem to be solved and large dimensional data can therefore be used without numerical problems. The second reason is that it easily allows nonlinear regression as an extension of the linear approach. The latter step is performed by replacing the inner product, $\langle x_i, x\rangle$, by a so-called kernel function: $K(x_i, x)$. If this function meets certain conditions[32] (Mercer's conditions), the kernel implicitly determines both a nonlinear mapping, $x \rightarrow \varphi(x)$, and the corresponding inner product $\varphi(x_i)^T\varphi(x)$. This leads to the following nonlinear regression function:

$$y_i = \sum_{i=1}^{n}\alpha_i K(x, x_i) + b \quad (6)$$

In principle, the nonlinear mapping can become infinite dimensional. For this case and if many input variables are present, solving this equation is particularly useful. However, for linear cases with not too many variables, eq 1 can also be used directly.

Finding the nonlinear mapping explicitly (i.e., without using kernels) can be very troublesome because for all input variables of the data, the specific mapping has to be known. This is especially difficult if the data are high dimensional such as spectra. Kernels typically used are the polynomial function, $\langle x_i,,x\rangle^d$, or the radial basis function, $\exp(-||x_i - x||^2/2\sigma)$, which is a Gaussian curve. As can be seen, each kernel is associated with a kernel-specific parameter. For the polynomial and RBF kernels, these parameters are the degree of the polynomial (d) and the width of the Gaussian function (σ), respectively. So instead of calculating a specific mapping for each dimension of the data, the problem comes down to selecting a proper kernel function and optimizing its specific parameter.

Important advantages of the LS-SVM approach are that it leads to a global solution that is often unique.[23] This is similar to PLS but an advantage over neural networks, for example. Furthermore, the dimension of the input data becomes irrelevant due to the inner product; therefore, nonlinear regression can be performed easily. This is a direct result from using the Lagrangian theory to solve the penalized cost function. As result, this approach requires finding the Lagrange multipliers that give a measure of the importance of a training object to the solution. A subselection of the most important training objects can be found by pruning the Lagrange multipliers. In this way, a sparse solution can be obtained. For SVMs, the deselection of irrelevant training objects follows inherently from the specific formulation of the cost function. Finally, the (nonlinear) LS-SVM regression model can be found by solving a set of linear equations, which is easy.

Note that, in contrast to the Lagrange multipliers, the choice of a kernel and its specific parameters together with γ do not follow from the optimization problem but have to be tuned by the user. These can be optimized by the use of Vapnik–Chervonenkis bounds, cross validation, an independent optimization set, or Bayesian learning.[15,23]

EXPERIMENTAL SECTION

Software. All calculations have been performed using Matlab. LS-SVM was performed using the Matlab/C toolbox.[33] For SVM, a Matlab toolbox was used.[34]

(32) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, MA, 2000.

Data. The data set used was originally described by Wülfert et al.[1] For the data set, NIR spectra were measured of ternary mixtures of ethanol, water, and 2-propanol. For these data, 19 different combinations of mole fractions are analyzed in a wavelength range of 850−1049 nm with a 1-nm resolution. Each mixture is measured at 30, 40, 50, 60, and 70 °C (±0.2 °C). It could be observed that measuring the spectra at different temperatures lead to nonlinear spectral variations; for this reason, relations between spectra from different temperatures cannot be made straightforward. The training set contains 13 mixtures per temperature while the independent test set contains 6 mixtures per temperature. The spectra have been baseline-corrected and mean-scaled. Using these data, in principle, two types of regression models can be made: global and local models. The global models are set up with training data from all the temperatures (65 objects) and are used to predict the (dimensionless) mole fractions at any temperature. In contrast, the local models are set up on basis of the training set from one temperature (13 objects), while predicting mole fractions of the test set for exclusively the same temperature. Obviously, to span the complete temperature range, as many local models as temperatures have to be made. Optimizing the models has been done by cross validation on the training set while the final prediction error has been based on the independent test set.

RESULTS

Optimizing the LS-SVM. As discussed above, the optimal LS-SVM model is obtained by finding the Lagrange multipliers that follow from minimizing the cost function using the Lagrange optimization procedure. However, minimizing the cost function is preceded by a definition of model parameters that influence the cost function: γ (the relative weight of the regression error) and d or σ (kernel parameters of the RBF or the polynomial kernel, respectively). In this paper, the optimal parameters are found from an intensive grid search. In practice, this numerical approach is the most common one used. The result of this grid search is an error-surface spanned by the model parameters. A robust model is obtained by selecting those parameters that give the lowest error in a smooth area.

In this paper, LS-SVM uses the often used Gaussian kernel (RBF). To find the optimal model parameters, for each of the three mixture compounds a grid search is performed on basis of 15-fold cross validation on the training set. The resulting LS-SVM models during cross validation were very similar concerning their support vectors. For the RBF parameter (σ), a series has been used of 0.1−2.5 with incremental steps of 0.1. For the γ, two different ranges have been used: 50−500 in steps of 50 and from 5×10^3 to 150×10^3 in steps of 2.5×10^3. In this way, parameter optimization was performed in different orders of magnitude. Because the grid search has been performed over just two parameters, a contour plot of the optimization error can be visualized easily (Figure 1). This is an advantage of LS-SVMs over SVMs in which three parameters have to be optimized. From Figure 1, it can be seen that the optimization error decreases on the diagonal from a high σ to a high γ. The optimal parameter settings can now be selected from (1) a smooth subarea with (2) a low prediction error. Similar error plots have been derived for the LS-SVM models when forecasting the mole fractions of water and 2-propanol. From these error plots, the following results are obtained for ethanol, water, and 2-propanol: $\gamma = 80\,500$, $\sigma = 0.6$; $\gamma = 73\,000$, $\sigma = 0.6$; $\gamma = 150\,500$, $\sigma = 0.7$, respectively. From these results, it appears that a relative large weight is given to the second part of the cost function due to the usage of a high γ. This means that emphasis has been put on obtaining low prediction errors while retaining possibly high weight coefficients. In principle, this again bears the risk of overfitting, but in this paper it is avoided by using cross validation for optimization and by invoking an independent test set to report the final results.

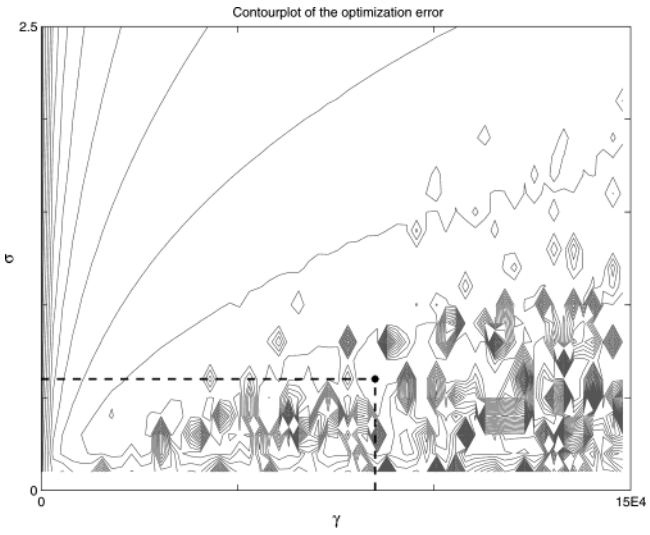

Figure 1. Contour plot of the optimization error for LS-SVM when optimizing the parameters γ and σ for the prediction of ethanol using a global model. The dot indicates the selected optimal settings.

MVC Results. When spectra are affected by unknown (nonlinear) interferences, the performance of MVC methods can deteriorate, and invalid conclusions might be drawn. Obviously, this is an unwanted situation that needs to be avoided. For an illustrative example in which NIR spectra are affected nonlinearly by different temperatures, the literature shows several approaches to make robust MVC methods based on PLS that will be reviewed shortly. In this case, robust means the ability to make accurate predictions irrespectively of the temperature.

The first approaches for this problem were to use either global or local models.[1] A global model has been made for all temperatures at once while local models have been set up for each possible temperature separately. Furthermore, attempts have been made to correct for the temperature influence by taking it into account explicitly as an extra variable, by performing robust variable selection, or by removing its influence with wavelets.[24] Variable selection has also been performed using simulated annealing.[26] In another PLS-based application, continuous piecewise direct standardization (CPDS) has been applied to remove nonlinear temperature effects.[25] The first nonlinear MVC approach used for this problem was a two-dimensional penalized signal

(33) Pelckmans, K.; Suykens, J. A. K.; Van Gestel, T.; De Brabanter, D.; Lukas, L.; Hamers, B.; De Moor, B.; Vandewalle, J. *LS-SVMlab: a Matlab/C Toolbox for Least Squares Support Vector Machines*; Internal Report 02-44, ESAT-SISTA; K. U. Leuven: Leuven, 2002. Available at http://www.esat.kuleuven.ac.be/sista/lssvmlab/.

(34) Gunn, S. R. *Support Vector Machines for Classification and Regression*; Technical Report; Image Speech and Intelligent Systems Research Group, University of Southampton: 1997. Available at http://www.isis.ecs.soton.ac.uk/isystems/kernel/.

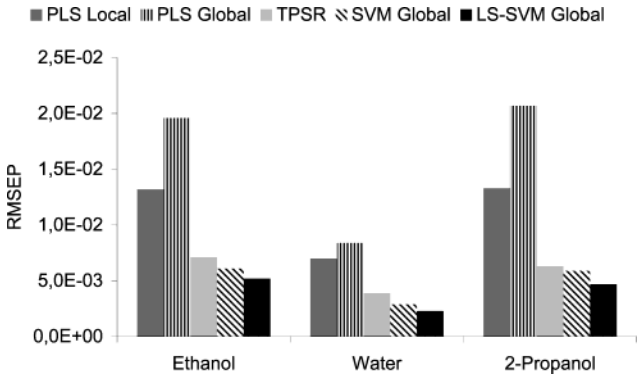

Figure 2. Calibration performances of different approaches from the literature together with the newly presented global model based on LS-SVM. The errors have been derived from an independent test set.[1] Note that these are dimensionless because the predicted variable represents a fraction.

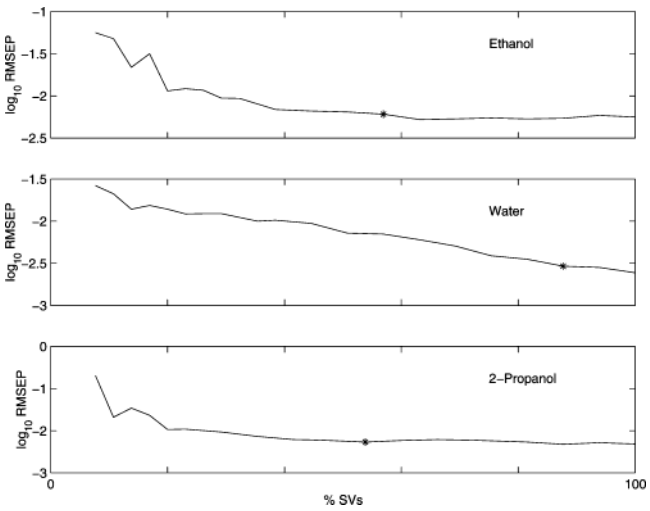

Figure 3. Logarithmic root-mean-square errors introduced when pruning the LS-SVM models for the prediction of ethanol, water, and 2-propanol. Indicated with an asterisk is the number of support vectors used for the pruned model. Note that the *y*-axis is dimensionless because the predicted variable represents a fraction.

regression method[28] (TPSR). TPSR uses a joined wavelength−temperature domain to determine the regression coefficients for an arbitrary temperature (global approach). In fact, this is an extension of the one-dimensional PSR that uses a newly formed basis of B-splines and forces the coefficients to vary smoothly with the wavelengths. Additionally, for a fair comparison with LS-SVMs, "traditional" SVMs have also been used to make a global model.[17]

Figure 2 shows a selection of the prediction results gathered directly from the literature. This selection contains the local PLS model (PLS local), which is the best PLS approach described. Furthermore, the global PLS method (PLS global) is included to enable a fair comparison with the other global methods. The optimal number of PLS factors for these models have been stated in the original papers and have been derived by cross validation. Finally, the nonlinear approaches have been included as well (global TPSR, SVM global, and LS-SVM global).

The figure shows that the nonlinear methods perform much better than the PLS-based models and that the LS-SVM global model outperforms all others. When comparing PLS local (best PLS approach) with LS-SVM global (the overall best approach), the latter has an RMSEP that is a factor of 2.6 lower. Except for better prediction ability, the (nonlinear) global approaches have the additional advantage that one model can be used for all temperatures. When mutually comparing the nonlinear modeling techniques, both SVM and LS-SVM outperform TPSR (leading to an RMSEP that is a factor of 1.1 and 1.3 lower, respectively). Finally, it can be seen that LS-SVM also performs better than its predecessor SVM, which is an additional advantage to its computational simplicity when compared with SVMs. Probably, a better performance is obtained because LS-SVM can be optimized much more accurately due to its computational simplicity (less parameters and much faster).

However, a possible advantage of SVMs over LS-SVMs is the fact that usually less support vectors than training objects are required in the model (sparseness). As it has been discussed above, LS-SVMs use all training objects in their final model; hence, no sparseness is obtained. However, model sparseness can be reinforced by using one of the existing pruning techniques applied to the Lagrange multipliers.[23,35] When using the approach of Suykens et al.,[23] first the LS-SVM model is built using all training objects. In the next step, those training objects are removed that are less relevant for the model (e.g., the objects corresponding to the lowest 5% of absolute Lagrange multiplier values). Next, this step is followed by a re-estimation of the model, after which again a certain number of training objects is removed. It should be stressed that re-estimation is a prerequisite to ensure that a new optimal model is found given a certain subset of (remaining) training objects. For this paper, the LS-SVM models have been pruned until the point where the prediction errors start to increase (Figure 3). After pruning, the number of LS-SVM support vectors were 37 (57%), 57 (88%), and 35 (54%) for ethanol, water, and 2-propanol, respectively. Pruning has been performed until an increase of the error of maximal 10% was achieved leading to an increase of 5×10^{-4}, 5×10^{-4}, and 6×10^{-4}, as compared to the full model. An exception for this is water. When pruning, its prediction error increases relatively much. However, these results are still comparable to or better than those of SVM. Furthermore, it appears that for ethanol and 2-propanol, SVM uses similar numbers of support vectors (60% and 55%, respectively); however, for water the number of SVM support vectors is much smaller (41.5%).[17] If the pruned LS-SVM model should also contain this number of support vectors, the prediction error increased with 7.1×10^{-3} (∼300%). The reason for this relatively high number of support vectors stems from the fact that the relative contributions of the training objects are more or less similarly relevant as compared to SVM. Therefore, removing only a few training objects can cause the prediction error to increase. Furthermore, Figure 5 also shows the important training objects to be more spread over the design.

Model Interpretation. The next step after having established a prediction model is to interpret the model. This can be done using the Lagrange multipliers. In this section, model interpretation is performed on the established SVM model,[17] the current LS-SVM model, and the pruned LS-SVM model. From Figure 4, it follows that the relative importance of the training objects are

(35) De Kruif, B. J.; De Vries, T. J. A. *IEEE Trans. Neural Networks* **2003**, *14*, 696−702.

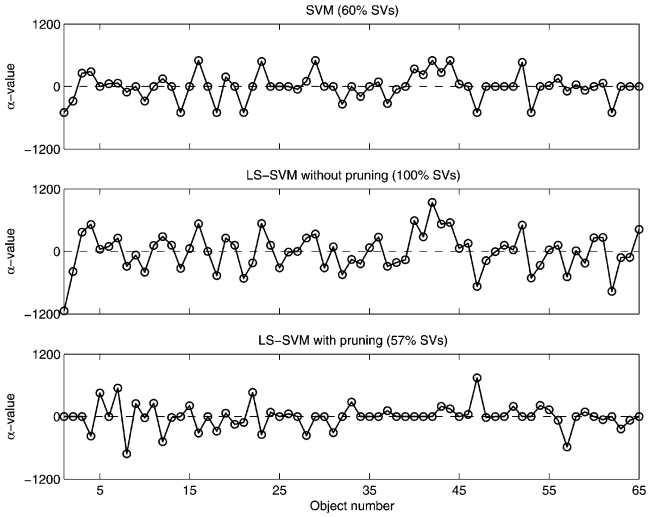

Figure 4. Values of the Lagrange multipliers from ethanol prediction models. The upper row shows the results for the SVM, while the middle and bottom rows show the Lagrange multipliers for LS-SVM and the pruned LS-SVM, respectively. Similar results are obtained for the prediction of water and 2-propanol.

very similar for SVM and LS-SVM (unpruned model) as is indicated by a correlation coefficient of 0.89 calculated for both sets of Lagrange multipliers. Training objects that are important for SVM are also important for LS-SVM. However, for LS-SVM also some objects are important that are irrelevant for SVM. For the pruned LS-SVM model, it appears that the relative importance of the training objects has changed and pruned series of Lagrange multipliers exhibit a correlation coefficient of only −0.12 as compared to the SVM model. Only for a part, the irrelevant SVM training objects coincide with irrelevant ones for the pruned LS-SVM (e.g., 15, 17, 25, or 65).

Figure 5 shows the location of the most important training objects in the mixture design (as obtained from Wülfert et al.[1]) for SVM, LS-SVM, and the pruned LS-SVM. In principle, for each model type, 15 mixture designs can be shown: for each of the three components to predict, five designs have been set up corresponding to the five different temperatures. Therefore, the importance of each mixture point has been obtained by taking the mean of the individual 15 mixture designs. The values shown therefore represent an overall value of 15 Lagrange multipliers. As a result, if some of these 15 Lagrange multipliers are set to zero, the overall importance can still be significant. The reason to aggregate the individual results for predicting different components into one figure is because the individual results were similar in sign and value.

It can be seen that SVM mostly uses training objects with a high mole fraction of ethanol and 2-propanol and low mole fractions of water. This can be explained from the fact that the NIR spectra of ethanol and 2-propanol are similar while the one from water deviates much more.[1] This means that it is more difficult to distinguish ethanol from 2-propanol than ethanol from water, for example. Hence, the reason objects 8, 13, 17, 18, and 19 are hardly important for the SVM model stems from the fact that the water contribution can be predicted better and the corresponding prediction error is smaller than ϵ. In the latter case, this means that the corresponding Lagrange multiplier value equals zero.[12] Due to the ϵ-insensitive error function, a very crisp distinction is created between important training objects and irrelevant ones. For the LS-SVM model, it appears that the importance of training points is spread over almost the whole design except for the mixtures with high water content. The prediction error of the latter samples is small; hence, their importance is still negligible. For the pruned LS-SVM, the important training objects are again much more spread over the design. It is shown that after removing the less relevant objects, objects 13 and 17 start to contribute more to the final model (this is possible because the importance of each mixture point is calculated from in total 15 points). This is caused by the fact that first many training objects with high water contents are removed because their contributions were only minor. This leaves only a few of these objects, which again become relatively important (after a few re-modeling steps) in order to enable the prediction of the water content.

So after having established the final LS-SVM model, the prediction time for new objects is related to the number of training objects in the data. Therefore, the prediction time for the standard LS-SVM model is higher than for the pruned LS-SVM model. This difference can play an important role if many training objects have been used and if prediction has to be performed on-line (i.e., fast). For those cases a pruned LS-SVM might be preferred over a regular LS-SVM. On the other hand, the unpruned LS-SVM allows a good interpretation of the model comparable to the traditional SVM.

DISCUSSION AND CONCLUSIONS

This paper proposes the use of LS-SVMs as a nonlinear calibration technique for solving ill-posed problems. Due to their

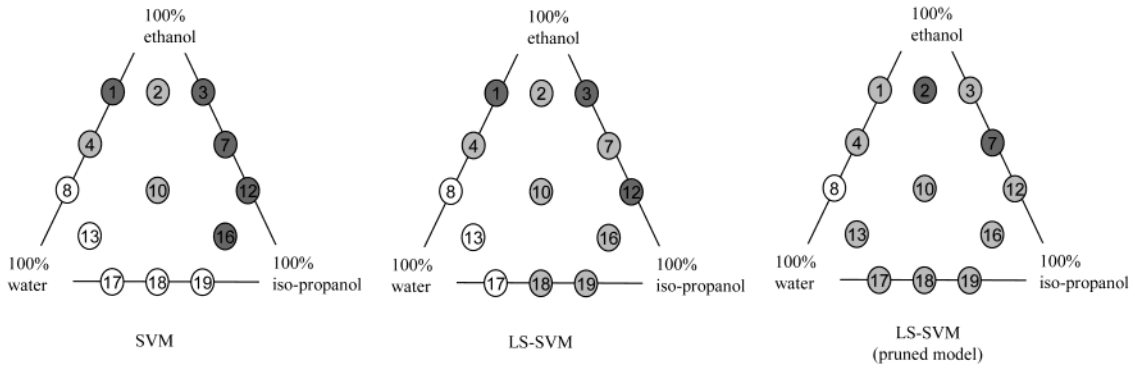

Figure 5. Relative importance of training objects for the three types of (LS-)SVM models. The white circles indicate an importance of less than 5%. The light gray and dark gray circles indicate an importance of 5−10% and more than 10%, respectively.

good prediction abilities, LS-SVMs are promising techniques to use in (analytical) laboratories as well as industries for solving nonlinear multivariate calibration problems. An important application is the estimation of the quality of products from indirect but fast and reliable measurements such as spectra. This improves the common approach of determining the quality parameters physically, which can be very time inefficient and allowing no on-line monitoring of product quality.

LS-SVMs most important advantages are that they lead to global (and often unique) nonlinear models that can be calculated easily. Using a well-known analytical chemical test case, this paper demonstrates the performance of LS-SVM. The test case shows to the difficult problem of relating temperature-affected NIR spectra to other characteristics of interest. Compared to the previously applied modeling methods to solve this problem, LS-SVMs perform best. Furthermore, strategies have been described regarding the optimization of the model, model pruning, and model interpretation. It appears that a pruned model can be obtained easily with a low prediction error. Additionally, the Lagrange multipliers can be used to interpret the importance of the training objects in the context of the considered analytical chemical problem.

Furthermore, although not applied in this paper, the extraction of the most informative regression features might also be a useful contribution to solve nonlinear MVC problems in a robust way. One way to obtain these results is to apply feature selection using optimization methods such as genetic algorithms, simulated annealing, or tabu search.[4] However, another solution for LS-SVM might be the use of a slightly different cost function leading to automatic feature selection. In the different cost function, the L_2 norm (sum of squared values) of the coefficients can be replaced by an L_1 norm (sum of absolute values). The feature of an L_1 norm is an inherent deselection of features because their coefficients are forced to zero. The L_1 norm alternative to RR is known as the least absolute shrinkage and selection operator method[36] (LASSO). For SVMs, this approach can be applied as well.[16,37]

Finally, performing kernel-based nonlinear mapping is shown to perform well but, thus far, it is not used to retrieve physicochemical information. Reconstructing the mapping according to Schölkopf et al.,[38] for example, and knowing what kind of mapping is preferred for specific features (e.g., spectral bands) can increase the knowledge of the problem. This in turn can give further directions to interpret and improve the results. Investigating this issue in combination with efficient feature selection, as discussed above, is one of the aspects our future research will focus on.

ACKNOWLEDGMENT

Paul Eilers (Leiden University Medical Center, Leiden, The Netherlands), Ton de Weijer and Erik Swierenga (Teijin Twaron, Arnhem, The Netherlands), Stan Gielen (Foundation for Neural Networks, Nijmegen, The Netherlands), and Sijmen de Jong (Unilever, Vlaardingen, The Netherlands) are thanked for constructive discussions on this work. This research was supported by the Dutch Technology Foundation STW (NCH4501).

Received for review December 20, 2003. Accepted March 30, 2004.

AC035522M

(36) Tibshirani, R. *J. R. Stat. Soc. B* **1996**, *58*, 267−288.
(37) Zhu, J.; Rosset, S.; Hastie, T.; Tibshirani, R. 1-Norm Support Vector Machines. *Neural Information Processing Systems Conference 2003*. Available at http://www-stat.stanford.edu/~hastie/pub.htm.
(38) Schölkopf, B.; Mika, S.; Burges, C. J. C.; Knirsch, P.; Müller, K.-R.; Rätsch, G.; Smola, A. *IEEE Trans. Neural Networks* **1999**, *10*, 1000−1017.