Determination of hydroxyl value of soybean polyol by least-squares support vector machines using FT-IR/HATR data

Marco Flôres Ferrão (UNISC) ferrao@unisc.br
Simone da Câmara Godoy (UFRGS) godoysc@iq.ufrgs.br
Annelise Engel Gerbase (UFRGS) agerbase@ufrgs.br
Cesar Mello (UNIFRAN) camello@unifran.br
Cesar Liberato Petzhold (UFRGS) petzhold@iq.ufrgs.br
Ronei Jesus Poppi (UNICAMP) ronei@iqm.unicamp.br

Summary

This paper presents the use of least-squares support vector machine (LS-SVM) for quantitative determination of hydroxyl value (OHV) of hydroxylated soybean oils by Fourier transform infrared spectroscopy with horizontal attenuated total reflection accessory (FT-IR/HATR). Calibration standards were prepared by the formic acid/hydrogen peroxide method and the OH values were determined by official method of AOCS Tx 1a-66, covering an analytical range of 23.66-195.04 mg of KOH/g per sample. A least-squares support vector machine (LS-SVM) calibration model for the prediction of hydroxyl value (OHV) was developed using the range 1805.1 to 649.9 cm^{-1}. Forty-two samples were used to model and twenty were used for external validation. Validation of the method was carried out by comparing the OHV of a series of hydroxylated soybean oil predicted by the LS-SVM model to the values obtained by the AOCS standard method. The performance in determining OHV of hydroxylated soybean oil samples by the LS-SVM has shown very good results. It was obtained a determination coefficient of 0.991 and RMSEP of 5.9061. This study shows that it is possible to determine OHV of hydroxylated soybean oil using LS-SVM by FT-IR/HATR spectra.

Keywords: *LS-SVM, infrared, hydroxylated soybean oil.*

1. Introduction

Recently neural networks as multilayer perceptrons and radial basis functions networks have been used in a wide range of fields, including control theory, signal processing and linear or nonlinear modeling (VAPNIK, 1998; SUYKENS, 2001). A promising methodology called support vector machines (SVM) (BURGES, 1998; SMOLA & SCHÖLKOPF, 2004) approaches to classification, nonlinear function and density estimation lead to convex optimization problems. In chemometrics applied, few and recent works had been used SVMs for the classification (BELOUSOV, VERZAKOV & VON FRESE, 2002 ; BRUDZEWSKI, OSOWSKI & MARKIEWICZ, 2004; LOZANO *et al.*, 2006) and quantification problems using spectra data set. Thissen *et al.* (2004) using for the determination of monomer masses during a copolymerization reaction by Raman spectra and quantification ethanol, water and iso-propanol in a ternary mixtures by near infrared (NIR) spectra. Chauchard *et al.* (2004) compared classical linear regression techniques and LS-SVM regression for the prediction of total acidity in fresh grapes and Cogdill *et al.* (2004) estimation of the physical wood properties using NIR spectroscopy.

Usually, the hydroxyl value (OHV) is determined by titration methods such as the American Oil Chemists' Society (AOCS, 1997) OHV determination (AOCS Cd 13-60) used in this

work. The hydroxyl value is expressed in mg of KOH per g of oil. This method is reliable and reproducible if carried out under standardized conditions, but it is time-consuming, labor-intensive, reasonably sensitive, largely dependent on the skills of the analyst, uses large amounts of sample and reagents, and some of them (pyridine, acetic anhydride) are hazardous and difficult to dispose off.

Similar problems were also observed in other chemical analyses of fats and oils based on titration methods. Therefore, spectroscopic methods are being increasingly used to replace wet chemical procedures. Infrared spectroscopy is one that has found increasing use due to its low cost, shorter time of analysis, non-destructiveness, small quantities of sample, in addition to accuracy and reliability when associated with chemometric methods (PARREIRA *et al.*, 2002; RUSSIN, VAN DE VOORT & SEDMAN, 2004). Moreover, FT-IR coupled with horizontal attenuated total reflectance (HATR) accessory simplifies many of the sample handling problems commonly associated with infrared analysis and is readily amenable to routine quality control applications (BORIN & POPPI, 2004).

In this work was proposed the multivariate regression model using least-squares support vector machine (LS-SVM) to determine hydroxyl value (OHV) of hydroxylated soybean oil by horizontal attenuated total reflection spectra.

2. Least-Squares Support Vector Machines (LS-SVM)

The Support Vector Machines (SVM) is a generalization of the Generalized Portrait algorithm developed in the sixties. However, the SVM, in its present form, was developed at AT&T Bell Laboratories in the nineties (THISSEN, 2003). Thus, SVM is a relatively nonlinear technique in the field of chemometrics and is employed basically in classification and multivariate calibration problems. The LS-SVM (SUYKENS & VANDERWALLE, 1999) is capable of dealing with linear and nonlinear multivariate calibration and resolves multivariate calibration problems in a relatively fast way. In LS-SVM a linear estimation is done in a kernel induced feature space ($y = w\phi(x) + b$). As in SVM, it is necessary to minimize a cost function (C) containing a penalized regression error, as follows:

$$C = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^{N} e_i^2 \tag{1}$$

such that

$$y_i = w^T \phi(x_i) + b + e_i \tag{2}$$

for all $i = 1, \ldots, N$; where $\phi(x_i)$ denotes the feature map.

The first part of this cost function is a weight decay, which is used to regularize weight sizes and penalize large weights. Due to this regularization, the weights converge to smaller values. Large weights deteriorate the generalization ability of the LS-SVM because they can cause excessive variance. The second part of equation 1 is the regression error for all training data. The parameter γ, which has to be optimized by the user, gives the relative weight of this part as compared to the first part. The restriction supplied by equation 2 gives the definition of the regression error.

Analyzing equation 1 and its restriction given by equation 2, it is possible to conclude that we have a typical problem of convex optimization (SUYKENS *et al.*, 2002), which can be solved

by using the Lagrange multipliers method (SUYKENS & VANDERWALLE, 1999), as follows:

$$L = \frac{1}{2}\|w\|^2 + \gamma\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N}\alpha_i\left\{w^T\phi(x_i)+b+e_i-y_i\right\} \tag{3}$$

Obtaining the optimum, that is, carrying out $\frac{\partial L(w,b,e,\alpha)}{\partial w}, \frac{\partial L(w,b,e,\alpha)}{\partial b}, \frac{\partial L(w,b,e,\alpha)}{\partial e}$, $\frac{\partial L(w,b,e,\alpha)}{\partial \alpha}$ and setting all partial first derivatives to zero, it obtains the weights that are linear combinations of the training data are obtained:

$$\frac{\partial L(w,b,e,\alpha)}{\partial w} = w - \sum_{i=1}^{N}\alpha_i\phi(x_i) = 0 \therefore w = \sum_{i=1}^{N}\alpha_i\phi(x_i) \tag{4}$$

$$\frac{\partial L(w,b,e,\alpha)}{\partial e} = \sum_{i=1}^{N}\gamma e - \alpha = 0 \therefore \alpha = \gamma e \tag{5}$$

then:

$$w = \sum_{i=1}^{N}\alpha_i\phi(x_i) = \sum_{i=1}^{N}\gamma e_i\phi(x_i) \tag{6}$$

where a positive definite kernel is used as follow:

$$K(x_i, x_j) = \phi(x_i)^T\phi(x_j) \tag{7}$$

An important result of this approach is that the weights (w) can be written as linear combinations of the Lagrange multipliers with the corresponding data training (x_i). Putting the result of equation 6 into the original regression line $(y = w\phi(x_i)+b)$, the following result is obtained:

$$y = \sum_{i=1}^{N}\alpha_i\,\phi(x_i)^T\,\phi(x_i)+b = \sum_{i=1}^{N}\alpha_i\left\langle\phi(x_i),\phi(x_j)\right\rangle + b \tag{8}$$

The α_i vector follows from solving a set of linear equations:

$$\mathbf{M}\alpha = \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{9}$$

where $\mathbf{M}$ is a square matrix as follow:

$$\mathbf{M} = \begin{bmatrix} \mathbf{K} + \dfrac{\mathbf{I}}{\gamma} & 1_N \\ 1_N^T & 0 \end{bmatrix} \tag{10}$$

where K denotes the kernel matrix with $i, j-th$ element $\mathbf{K} = (x_i, x_j) = \phi(x_i)^T\phi(x_j)$ and $\mathbf{I}$ denotes the identity matrix $\mathbf{N}\,x\,\mathbf{N}$, $1_N = [1 \quad 1 \quad 1 \quad \cdots \quad 1]^T$. Hence the solution is given by:

$$\alpha = \mathbf{M}^{-1} \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{11}$$

As can be seen from equations 10 and 11, usually all Lagrange multipliers (the support vectors) are nonzero, which means that all training objects contribute to the solution. In contrast with standard SVMs the LS-SVM solution is usually not sparse. However, as described in Suykens *et al.* (2002) a sparse solution can be easily achieved by pruning or reduction techniques. Depending on the number of training data set either direct solvers can be used or iterative solver such as conjugate gradients methods (for larges data sets), in both cases with numerically reliable methods. In applications involving nonlinear regression it is enough to change the inner product $\langle \phi(x_i), \phi(x_j) \rangle$ of equation 8 by a kernel function, $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$. If this kernel function meets Mercer's condition (MERCER, 1909) the kernel implicitly determines both a nonlinear mapping, $x \to \phi(x)$ and the corresponding inner product $\phi(x_i)^T \phi(x)$. This leads to the following nonlinear regression function:

$$y_i = \sum_{i=1}^{N} \alpha_i K(x_i, x_j) + b \tag{12}$$

The attainment of the kernel function is cumbersome and it will depend on each case. However, the kernel function more used is the Radial Basis Function (RBF), $\exp\left(-\|x_i - x\|^2 / 2\sigma\right)$, a simple Gaussian function, and polynomial functions $\langle x_i, x_j \rangle^d$, where *d* is the polynomial degree and σ is the width of the Gaussian function, which should be optimized by the user, to obtain the support vector. We stressed that it is very important to do a careful model selection of the tuning parameters, in combination with the regularization constant γ, in order to achieve a good generalization model.

3. Experimental

3.1. Chemicals:

Refined soybean oil was purchase from was supplied by CBM Ind. Com. Distrib. Ltda (Cachoeirinha, RS, Brazil). Formic acid and ethyl ether were purchased from Synth. Hydrogen peroxide solution 30%, sodium chloride, sodium carbonate, sodium hydrogensulfite, sodium sulfate anhydrous were purchase from Nuclear. All chemicals are analytical grade and were used without further purification.

3.2. Calibration Standards:

Soybean polyols were synthesized following the method described below and were used as calibration standards. Depending on the time of reaction, soybean polyols with different OH functionality were obtained. The acid number (AN) and the OHV of the soybean polyols were determined by the AOCS standard method Cd 3a-63 (AOCS, 1980) and Cd 13–60 (AOCS, 1997), respectively. The OHV cover a range of 23.66-195,04 mg. KOH per g of sample.

3.3. Instrumentation:

A Nicolet Magna 550 FT-IR spectrophotometer with a 4 cm^{-1} resolution and 16 scans was used for the measurement of soybean polyols. The duplicate spectra were recorded by applying the soybean polyol sample on the surface of a Pike horizontal attenuated total reflectance (HATR) sample-handling accessory with ZnSe crystal.

3.4. Multivariate modeling:

The average specters for each samples had been gotten using the two replicates obtained. Data were treated with multiplicative scatter correction (MSC) technique before further multivariate analysis.

Calibration: The LS-SVMlab (Matlab/C Toolbox for Least Squares Support Vector Machines, SUYKENS *et al*., 2002) was used for developed multivariate model. The program was run on an IBM-compatible Intel Pentium 4 CPU 3.00 GHz and 1 Gbytes RAM microcomputer. FT-IR/HATR spectra of the 42 soybean polyol samples were used. The samples presenting extreme OH values were included in the calibration set. Cross-validation following the leave-one-out procedure was performed during the validation step in order to define the optimum number of factors that should be kept in the model and to detect any outliers.

Validation: Spectra of twenty soybean polyol samples were used for the validation of the multivariate regression models. To evaluate the error of calibration model, the root mean square error was used, calculated by equation 13:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \qquad (13)$$

where n is the number of spectra, y_i and $\hat{y}_i$ are the values determined by AOCS standard method and those predicted by LS-SVM model, respectively, in the calibration set (RMSEC) or external validation set (RMSEP).

4. Results and discussion

Low-molecular-weight liquid epoxy polyol esters or ethers from vegetable oils can be employed as polyols in polyurethane formulation. Usually, hydroxyl groups have been introduced through a two-step synthesis involving firstly the epoxidation of the unsaturated sites with formic acid and hydrogen peroxide, followed by epoxy ring opening with mono or polyfunctional alcohols, amino alcohols, or acids. Depending on the reaction conditions, polyols with high OH functionality (complete reaction) or epoxy polyol esters with remaining epoxy groups (partial conversion) are obtained.

In this work, epoxy polyol esters were prepared by “one-step” synthesis using the formic acid/H_2O_2 system. The hydroxylation reaction was carried out at constant temperature, 65^{o}C, and by increasing the reaction time it was possible to prepare soy polyols with different OH functionality, which were used as calibration standards and for external validation samples.

The spectra of the 62 samples were pre-processed by multiplicative scatter correction (MSC), aiming at correcting the baseline deviation between the spectra.

In Figure 1 the optimized surface result for LS-SVM model, using the spectra calibration set, is shown. The γ and σ^2 parameters were a manageable task, similar to the process employed to select the number of latent variables for PLS models, but in this case for a two-dimensional problem. Was used the cross-validation procedure for to determine the RMSECV.

In Table 1 compares the coefficient of correlation (R^2_{cal}) for the calibration model, RMSECV and RMSEP for the different γ and σ^2 parameter combinations. When γ = 200 was used and increase σ^2 the RMSE values are optimized even σ^2 = 8000. When σ^2 = 8000 was used and γ

was increase, the RMSECV value are not optimized and increase of the risk of over-fitting. It's possible to observe that when increase the σ^2 value toward infinity, the RMSECV value tends to a minimum. However, this would likely lead to an over-fitted calibration.

The graph between measured and HATR predict OH is presented in Figure 2 for the 42 calibration and 20 prediction spectra samples. The correlation coefficients (R^2) between three PLS and LS-SVM models for OH determining were found to be the best. But it's possible observed in LS-SVM model presents better predict ability for the samples which low or upper values, which indicated that the LS-SVM calibration model could be used for this case with minor errors.

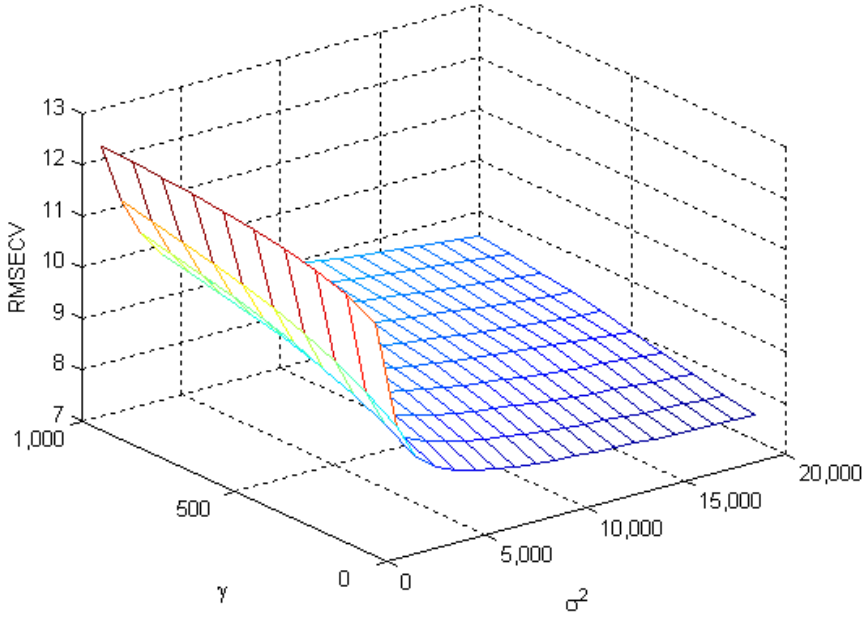

Figure 1 – Parameter optimization response surface for LS-SVM model.

γ	σ^2	RMSECV	R^2_{cal}	RMSEP
200	2000	9.5559	0.9939	7.3922
200	5000	8.2869	0.9922	6.0583
200	8000	8.0147	0.9907	5.9061
200	10000	7.9002	0.9901	5.9121
500	8000	8.5753	0.9928	5.8377
800	8000	8.9285	0.9921	5.7564

Table 1 – Performance comparison results for different LS-SVM training parameters.

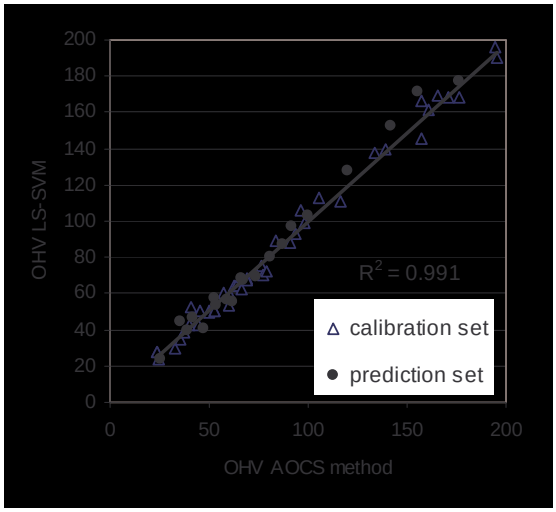

Figure 2 – Calibration and prediction plot of OHV for LS-SVM model.

5. Conclusions

This paper proposes the application of LS-SVMs to quantifying hydroxyl value of hydroxylated soybean oil samples. The results demonstrate a better prediction ability of the LS-SVM techniques to determine OHV in hydroxylated soybean oil samples. OHV determination was carried out in 2-3 min per sample, which is a major improvement over conventional wet chemical methods. Another important advantages was the capacity of the LS-SVM to predict extreme values in the samples, when the PLS models present major errors.

Finally, LS-SVMs are promising techniques to use for estimation of the quality of the products from indirect but fast and reliable measurements such as FT-IR/HATR spectra.

Acknowledgements

This work was supported by the Brazilian National Research Council (CNPq).

References

AL-ALAWI, A. & VAN DE VOORT, F.R. *New method for the quantitative determination of free fatty acids in oil by FTIR spectroscopy.* Journal of the American Oil Chemists' Society. Vol.81, p. 441-446, 2004.

AMERICAN OIL CHEMISTS' SOCIETY. *Official Methods and Recommended Practices of de American Oil Chemists' Society*, 4th edn, Champaing, 1980.

AMERICAN OIL CHEMISTS' SOCIETY. *Official Methods and Recommended Practices of de American Oil Chemists' Society*, 4th edn, Champaing, 1993, revised 1997.

BELOUSOV, A.I.; VERZAKOV, S.A. & VON FRESE, J. *Applicational aspects of support vector machines.* Journal of Chemometrics. Vol.16, p.482-489, 2002.

BORIN, A. & POPPI, R.J. *Multivariate Quality Control of Lubricating Oils Using Fourier Transform Infrared Spectroscopy.* Journal of the Brazilian Chemical Society. Vol.15 , n.4, p. 570-576, 2004.

BRUDZEWSKI, K.; OSOWSKI, S. & MARKIEWICZ, T. *Classification of milk by means of an electronic nose and SVM neural network.* Sensors and Actuators . B , Chemical. Vol. 98, p.291-298, 2004.

BURGES, C.J.C. *A tutorial on support vector machine for pattern recognition*. Data Mining and Knowledge Discovery. Vol.2, n.2, p.121-167, 1998.

CHAUCHARD, F.; COGDILL, R.; ROUSSEL, S.; ROGER, J.M. & BELLON-MAUREL, V. *Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes.* Chemometrics and Intelligent Laboratory Systems. Vol.71, 141-150, 2004.

COGDILL, R.P.; SCHIMLECK, L.R.; JONES, P.D.; PETER, G.F.; DANIELS, R.F. & CLARK, A. *Estimation of the physical wood properties of Pinus taeda L. radial strips using least squares support vector machines*. Journal of Near Infrared Spectroscopy. Vol.12, p. 263-270, 2004.

LOZANO, J.; SANTOS, J.P.; ALEIXANDRE, M.; SAYAGO, I.; GUTIERREZ, J. & HORRILLO, M.C. *Identification of typical wine aromas by means of an electronic nose*. IEEE Sensors Journal. Vol.6, n.1, p. 173-178, 2006.

MERCER, J. *Functions of positive and negative type and their connection with the theory of integral equations*. Philosophical Transactions of the Royal Society of London. Series A. Vol.209, p. 415-446, 1909.

PARREIRA, T.F.; FERREIRA, M.M.C.; SALES, H.J.S. & ALMEIDA, W.B. *Quantitative Determination of Epoxidized Soybean Oil Using Near Infrared Spectroscopy and Multivariate Calibration*. Applied Spectroscopy. Vol.56, p.1607-1614, 2002.

SMOLA, A.J. & SCHÖLKOPF, B. *A tutorial on support vector regression*. Statistics and Computing. Vol.14, p.199-222, 2004.

SUYKENS, J.A.K & VANDERWALLE, J. *Least-Square support vector machine classifiers*. Neural Processing Letters. Vol.9, p.293-300, 1999.

SUYKENS, J.A.K. *Support vector machines: A nonlinear modeling and control perspective*. European Journal of Control. Vol.7, p.311-327, 2001.

SUYKENS, J.A.K.; VAN GESTEL, T.; DE BRABANTER, J.; DE MOOR, B. & VANDEWALLE, J. *Least-Squares Support Vector Machines*. World Scientific, Singapore, 2002.

THISSEN, U.; VAN BRAKEL, R.; DE WEIJER, A.P.; MELSSEN, W.J. & BUYDENS, L.M.C. *Using support vector machines for time series prediction.* Chemometrics and Intelligent Laboratory Systems. Vol.69, p.35-49, 2003.

THISSEN, U.; PEPERS, M.; ÜSTÜN, B.; MELSSEN, W.J. & BUYDENS, L.M.C. *Comparing support vector machines to PLS for spectral regression applications.* Chemometrics and Intelligent Laboratory Systems. Vol.73, p.169-179, 2004.

VAPNIK, V. *Nonlinear Modeling: Advanced Black-Box Techniques*, ed. SUYKENS, J.A.K. & VANDEWALLE, J., Kluwer Academic Publishers, Boston, 1998.