Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied…

Article *in* Food Chemistry · April 2014

DOI: 10.1016/j.foodchem.2013.10.020 · Source: PubMed

CITATIONS
37

READS
337

3 authors:

Olivier Devos
Université des Sciences et Technologies de Li…
56 PUBLICATIONS **392** CITATIONS
SEE PROFILE

Gerard Downey
TEAGASC - The Agriculture and Food Develop…
174 PUBLICATIONS **5,090** CITATIONS
SEE PROFILE

Ludovic Duponchel
University of Lille Nord de France
83 PUBLICATIONS **1,046** CITATIONS
SEE PROFILE

Food Chemistry 148 (2014) 124–130

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Analytical Methods

Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils

Olivier Devos [a,*], Gerard Downey [b], Ludovic Duponchel [a]

[a] Laboratoire de Spectrochimie Infrarouge et Raman (LASIR CNRS UMR 8516), Université de Lille 1 Sciences et Technologies, Bât. C5, 59655 Villeneuve d'Ascq, France
[b] Teagasc Food Research Centre Ashtown, Ashtown, Dublin 15, Ireland

ARTICLE INFO

Article history:
Received 18 February 2013
Received in revised form 19 September 2013
Accepted 2 October 2013
Available online 14 October 2013

Keywords:
Genetic algorithm
Spectral pre-processing
Parameter optimisation
Classification
Support vector machines
Infrared spectroscopy

ABSTRACT

Classification is an important task in chemometrics. For several years now, support vector machines (SVMs) have proven to be powerful for infrared spectral data classification. However such methods require optimisation of parameters in order to control the risk of overfitting and the complexity of the boundary. Furthermore, it is established that the prediction ability of classification models can be improved using pre-processing in order to remove unwanted variance in the spectra. In this paper we propose a new methodology based on genetic algorithm (GA) for the simultaneous optimisation of SVM parameters and pre-processing (GENOPT-SVM). The method has been tested for the discrimination of the geographical origin of Italian olive oil (Ligurian and non-Ligurian) on the basis of near infrared (NIR) or mid infrared (FTIR) spectra. Different classification models (PLS-DA, SVM with mean centre data, GENOPT-SVM) have been tested and statistically compared using McNemar's statistical test. For the two datasets, SVM with optimised pre-processing give models with higher accuracy than the one obtained with PLS-DA on pre-processed data. In the case of the NIR dataset, most of this accuracy improvement (86.3% compared with 82.8% for PLS-DA) occurred using only a single pre-processing step. For the FTIR dataset, three optimised pre-processing steps are required to obtain SVM model with significant accuracy improvement (82.2%) compared to the one obtained with PLS-DA (78.6%). Furthermore, this study demonstrates that even SVM models have to be developed on the basis of well-corrected spectral data in order to obtain higher classification rates.

1. Introduction

Vibrational spectroscopy techniques (near infrared, mid-infrared and Raman) are rapid, non-destructive and generally non-invasive measurement methods. While food applications of Raman spectroscopy are now emerging, the infrared methods are already widely used in the agriculture, food and pharmaceutical industries for proximate analysis and quality control. In general, vibrational spectra contain compositional data which can be extracted using multivariate mathematical tools to yield quantitative (using regression models e.g. partial least squares regression, PLSR) or qualitative information (using classification e.g. partial least squares discriminant analysis, PLS-DA) or class-modelling (e.g. soft independent modelling of class analogy, SIMCA) solutions. It is common practice to apply a data pre-processing step to raw spectral data prior to modelling; this is because of the fact that, in addition to chemical information, infrared spectra (especially NIR) contain random and systematic interferences from other sources (i.e. noise, stray light, light scatter, detector non-linearities, temperature variations, etc.) which have the potential to degrade model performance and therefore should be removed. Typical pre-processing methods include multiplicative scatter correction (MSC), standard normal variate (SNV), 1st and 2nd derivatives (1Der and 2Der). Experience has shown that the use of one or more of these transformations can improve classification accuracy in the case of qualitative analysis and increase prediction accuracy of quantitative models.

With particular regard to qualitative analysis, many methods exist for sample classification based on spectroscopic data (Balabin, Safieva, & Lomakina, 2010; Berrueta, Alonso-Salces, & Heberger, 2007). Very often the final choice of a classification algorithm depends on the structure of data being studied but the final selection criterion remains first and foremost the prediction performance obtained with any model. Support vector machines (SVMs) belong to a new generation of learning algorithms used for classification and regression tasks (Cristianni

* Corresponding author. Tel.: +33 320 434 748; fax: +33 320 436 755.
E-mail address: olivier.devos@univ-lille1.fr (O. Devos).

0308-8146/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.foodchem.2013.10.020

& Shawe-Taylor, 2000; Schölpkopf & Smola, 2002). SVMs have been introduced as chemometric tools quite recently but have been successfully applied to mid and near infrared classification tasks such as material identification and food discrimination (Balabin et al., 2010; Xu, Zomer, & Brereton, 2006). In the case of classification, SVM simultaneously minimises the empirical classification error and maximises the inter-class geometric margin. One of the major features of SVM models is that they can operate in a kernel-induced feature space allowing for non-linear modelling. Other advantages are that there is a unique, optimal solution when the model parameters are fixed and that good generalisation performance is obtained even with relatively small datasets. It has been reported that classification models obtained using this technique are robust and less subject to the curse of dimensionality and over-fitting, properties that need to be borne in mind when dealing with spectroscopic data.

Nevertheless this method has some drawbacks, the most significant of which is the need to optimise several meta–parameters which are not easy to tune (model selection) and which is usually done by minimising either an estimate of the error or some related performance measure. Cross-validation over a grid search is a simple and universal strategy (Devos, Ruckebusch, Durand, Duponchel, & Huvenne, 2009) but can be computationally expensive; furthermore model accuracy is subject to the chosen granularity of the grid. A stepwise approach can also be used: initial SVM meta–parameters are iteratively updated in the direction of error decrease until the optimum values are found. The direction of the decrease can be determined using a gradient (Chapelle, Vapnik, Bousquet, & Mukherjee, 2002; Tseng & Yun, 2010) or gradient-free approach (Momma & Bennett, 2002).

However, as currently practised, SVM meta parameters and other parameters such as those involved in data pre-processing require separate steps for their individual optimisation; this is time-consuming and may not, in fact, produce optimal models so this paper proposes a methodology to simultaneously optimise both the pre-processing and SVM meta–parameters for a classification task. The strategy adopted in this paper is based on a Parallel Genetic Algorithm (called GENOPT-SVM) with chromosomes coding both pre-processing and SVM meta–parameters. It is evaluated on two separate spectral datasets (one near infrared and the other mid-infrared) collected from a single set of olive oil samples of known provenance in order to predict the geographic origin of the oils from their near or mid infrared spectra. This work also explored whether pre-processing was as useful for SVM models as it has been shown to be for other classification methods (e.g. PLS-DA).

2. Theory

2.1. SVM classification

The theory of SVM has been extensively described in literature (Cristianni & Shawe-Taylor, 2000; Schölpkopf & Smola, 2002) and only a brief description of the concept as regards classification will be given here. In considering a binary classification problem, the objective is to predict for all objects their membership of a class y (−1, +1) from m dimensional input data represented by a vector written $X = (x_1, x_2, \ldots, x_m)$. In the case of spectra, m represents the number of wavelengths. The class prediction first requires training on a data set containing the spectra corresponding to n objects or samples with known class, that is to say n (X_i, y_i) values where X_i is the ith object of the training set.

The idea behind the linear SVM algorithm is the search for the "best" hyperplane (called the optimal hyperplane) separating the data classes. This hyperplane can be described by a single linear equation (Eq. (1)):

$$w.X_i + b = 0 \tag{1}$$

where w (normal vector to the hyperplane) and b (offset) are calculated during SVM training. In the training stage, SVM tries to find the hyperplane (see Fig. 1) which minimises the classification error while simultaneously maximising the shortest distances from this hyperplane to the closest training samples of each class (d_+ for class (+1), d_- for class (−1)). The distance d_+ and d_-, equal to $\frac{1}{\|w\|}$, define the margin associated with the separating hyperplane. Optimisation of this margin is obtained by solving the constrained quadratic optimisation problem (Eq. (2)):

$$\begin{aligned} &\text{minimise} \quad (\|w\|^2 + C\sum_{i=1}^{n}\zeta_i) \\ &\text{subject to } \zeta_i + y_i(w.X_i + b) - 1 \geqslant 0 \text{ with } \zeta_i \geqslant 0 \end{aligned} \tag{2}$$

SVM tries to maximise the margin while keeping the classification error as low as possible. The classification error here is represented by the distance ξ_i of the misclassified sample i and the corresponding margin hyperplane. C called the regularisation meta–parameter, controls the trade-off between the two conflicting objectives: when C is small, margin maximisation is emphasised whereas when C is large, error minimisation is predominant.

According to optimisation theory (Bazaraa, Sherali, & Shetty, 2006), the Lagrangian dual formulation, which expresses the importance of each example in the training set, can be used to solve this problem and the optimal hyperplane may be finally expressed as a linear combination of the training observations (Eq. (3)).

$$f(X) = w.X + b = \sum_{i=1}^{n} y_i \alpha_i X_i X + b \text{ with } \alpha_i \geqslant 0 \tag{3}$$

where α_i corresponds to a coefficient, called a Lagrange multiplier, associated with each object. Only the samples close to the boundary (with $\alpha_i \neq 0$, called support vectors) are necessary to calculate the decision boundary function.

SVM classification methodology can be extended to nonlinear classification using a mapping function. Data are first projected, with the use of a mapping function φ, on a higher dimensional feature space; a linear SVM is then applied in this feature space where, ideally, the data can be linearly separable. In fact, in the dual representation, the explicit mapping is not required – only the inner products between two objects $\varphi(X_1).\varphi(X_2)$ is used and this product is called a kernel. Among existing kernel functions, the radial basis function (RBF) kernel is the most widely used as almost any boundary shape can be obtained with this kernel and

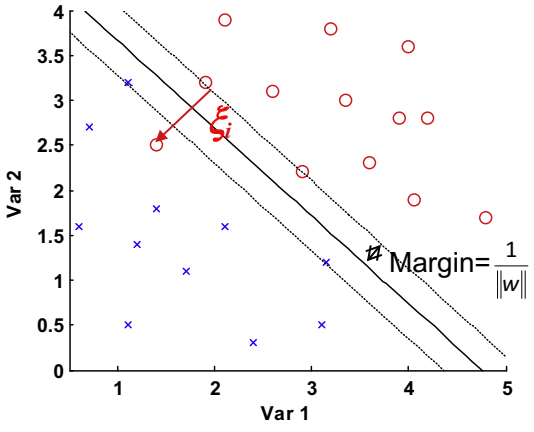

Fig. 1. Example of linear SVM classification for 2 variables.

models with good performance are generally obtained. The RBF kernel function is given in Eq. (4) where φ is the mapping function, X_1 and X_2 two objects and G is related to the kernel width meta–parameter.

$$K(X_1, X_2) = \varphi(X_1).\varphi(X_2) = exp(-G\|X_1 - X_2\|) \quad (4)$$

Thus, the use of SVM involves tuning two meta–parameters beforehand: the regularisation parameter C and the kernel parameter G for the RBF kernel parameter. This is a key step in SVM as their combined values determine the boundary complexity and of course the observed classification rate.

2.2. Genetic algorithms

Genetic algorithms (GAs) are stochastic methods which imitate natural evolution (Holland, 1992) and they have been successfully used for global search and optimisation problems (Mitchell, 1998).

In the case of optimisation, GAs are based on a population of individuals all coded by their own chromosome which represent possible solutions. Each individual is tested in terms of performance (objective function) for the given task. The best individuals are used to build the next generation. The successive steps used are the following: initialisation of the first generation (genesis), evaluation of individuals, selection of the best individuals and creation of the next generation using specific genetic operators (crossover, mutation...). This process is then repeated until a termination condition has been reached. A more complete description can be found in Goldberg (1989) and Talbi (2009).

2.3. GENOPT-SVM and pre-processing selection

The chromosome consists of three parts coding the two meta–parameters C and G and the pre-processing to apply to the data. A binary representation is used for the whole chromosome while both C and G are coded in bit string. The number of digits in the bit string and the range of the parameter (minimum and maximum values) are fixed by the user. The binary string is first converted into a decimal value and finally the values for C and G can be retrieved using the following equation:

$$\text{Parameter value} = min_p + \frac{max_p - min_p}{2^l - 1} \times d \quad (5)$$

where min_p and max_p are, respectively, the minimum and maximum value of the parameters, l is the length of the bit string and d is the decimal value of the bit string.

Each pre-processing method is coded on several bits depending on the number of pre-processing functions we want to test. For example, a five bits string can code up to 2^5 (32) pre-processing methods; furthermore, p pre-processing methods can be consecutively applied. In this case, all pre-processing bit strings are concatenated. During the fitness evaluation, all pre-processing methods are applied to the full spectra in the same order as they appear in the chromosome (from left to right). The different steps of GENOPT-SVM are presented in Fig. 2. During the initialisation step, each bit value is assigned randomly.

For a classification task, the fitness value is classically the classification error rate or the percentage of correct classification as determined by cross-validation.

To cover an important part of the solutions space, a large population and a high number of generations are needed before reaching convergence; this requirement results in a considerable computation time. To speed up the calculation, a parallel GA (PGA) is used instead of a single GA (SGA); details of the strategy and implementation we have developed can be found in Devos and Duponchel (2011).

2.4. Fitness evaluation and comparison of classification models

In this work, fitness value is estimated using the classification error of SVM obtained from the training dataset; the aim of the GA algorithm is therefore to minimise this value. Either a k-fold or a leave-one-out cross-validation can be used to calculate the classification error; generally a k-fold cross-validation is preferred due to lower computation time required.

At the end of GA optimisation, an independent test set is used in order to compare the different classification models. The percentage of total correct classification, sensitivity and selectivity for classification models are then calculated. For a binary classification, the sensitivity and selectivity are the percentage of correct classification for the positive or negative classes, respectively. McNemar's test (McNemar, 1947; Roggo, Duponchel, Ruckebusch, & Huvenne, 2003), a paired version of the χ^2 test, is used in order to compare two classification models in term of accuracy. It relies on calculation of the statistics presented here (Eq. (6)) with a continuity correction term

$$\text{Mc Nemar's value} = \frac{(|n_{o1} - n_{10}| - 1)^2}{n_{o1} + n_{10}} \quad (6)$$

where n_{01} is the number of samples misclassified by the first model only and n_{10} is the number of samples misclassified by the second model only. To be able to apply McNemar's test the sum of n_{01} and n_{10} must be superior to 20 which is the case in the results presented here. This statistic is distributed (approximately) as χ^2 with 1 degree of freedom, and the critical value for a 5% significance level is 3.841. If the McNemar's value is greater than this critical value, the null hypothesis (same error rate for the two models) is rejected and the two models are considered significantly different.

3. Experimental

Detailed information about the samples and instrumentation can be found in (Woodcock, Downey, & O'Donnell, 2008) and (Hennessy, Downey, & O'Donnell, 2009). The classification task was to discriminate olive oil coming from Liguria (a coastal region of north-western Italy) and olive oil samples from other regions of Italy on the basis of NIR or MIR spectra. Extra virgin olive oil samples (210 Ligurian and 700 non-Ligurian) were collected by marketing or regulatory bodies over three consecutive harvests (2005, 2006 and 2007) so their authenticity can be accepted with a high degree of confidence. In this application, model sensitivity refers to the percentage of Ligurian samples which are correctly classified by the Ligurian model while selectivity is the percentage of non-Ligurian samples which are correctly rejected by the same model.

For NIR measurement (Woodcock et al., 2008), olive oil samples were maintained at 30 °C during spectral acquisition. Transflectance spectra (1100–2498 nm) were collected using a camlock cell and a gold-plated reflector (0.1 mm sample thickness; part 99213) on a scanning spectrophotometer (NIRSystems 6500, NIRSystems Inc., Silver Spring MD). The NIR spectra are presented in Fig. S1.

In the case of MIR measurements, olive oil samples were maintained during 1 h at 25 °C prior to spectral collection (Hennessy et al., 2009). FT-IR spectra (600–4000 cm^{-1}) were collected on a Bio-Rad Excalibur series FTS 3000 spectrometer (Analytica Ltd., Dublin, Ireland). Samples were applied to an in-compartment benchmark attenuated total reflectance (ATR) trough plate using a 45° germanium crystal with 11 internal reflections (Specac Ltd., Kent, UK). The spectral zone between 2250 and 2400 cm^{-1} has been removed since the absorption of atmospheric carbon dioxide is observed in this region. Furthermore the ends of the spectral range containing mainly noise have been removed too. Finally only

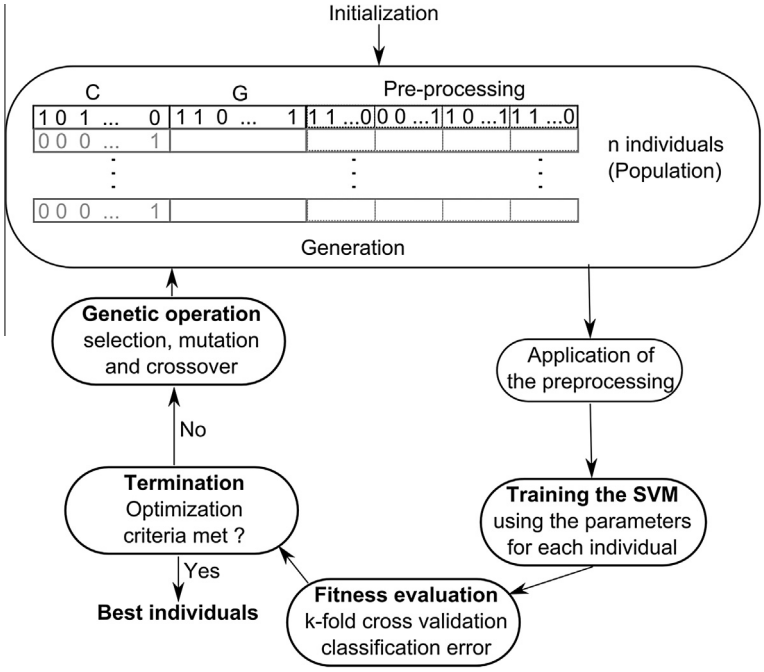

Fig. 2. Scheme of GENOPT-SVM for parameter optimisation and pre-processing selection.

the spectral zones from 3000 to 2400 cm^{-1} and from 2250 to 700 cm^{-1} have been used (see Hennessy et al., 2009). MIR spectra are presented in Fig. S2.

Separation of the sample collection into a calibration set (280 samples of which 140 originated in Liguria) and a validation set (630 samples of which 70 originated in Liguria) produced identical such sets for the NIR and MIR datasets.

In the case of high dimensional data the use of a dimension reduction method before SVM classification may reduce over fitting and increase the predictive ability especially when the number of samples in the training set is small. For the two datasets presented here the SVM results obtained using the whole spectra or using only the first principal components are similar both for calibration and validation. Therefore, in the following, SVM has been applied directly on the whole spectra.

3.1. GA configurations

Many parameters need to be tuned before using GENOPT-SVM. Fortunately, in our case most of them affect mainly the computation time and not the final solutions. Table S1 presents parameters used in this article and values associated with them which allow a small classification error after a few generations without premature convergence.

Since computation time is not a limiting problem due to the use of a computing cluster with parallel GAs, a large population with 256 individuals has been chosen. The pre-processing methods are coded on 5 bits representing 32 different procedures with various settings which represent those most commonly used with infrared spectra (Rinnan, Berg, & Engelsen, 2009) (Table S2). Before optimisation, the user can choose to apply only one pre-processing (5 bits) or a combination of n (from 2 to 4) successive pre-processing steps (coded on $n \times 5$ bits). In the case of combined pre-processing before evaluation, the population is checked to establish whether the same pre-processing occurs more than once for any given individual. In such cases, it is discarded and replaced by a new one.

The fitness evaluation is based on the classification error calculated by 10-fold cross-validation. Forty bits are used to code both C and G while their ranges have been fixed to those classically used for a grid search approach.

At each generation, 50% of the parents (128 individuals) are selected by the stochastic universal sampling method using a single point cross-over in order to generate children; a mutation probability set at 0.5% is also used. The new population is then composed of the 128 parents and the 128 children with the best fitness (elitist reinsertion). In this way, this new generation is more adapted to its environment than the previous one.

Eventually the optimisation stops when 50% of the chromosomes of the final generation share exactly the same bit string. Less than 15 generations were required to obtained convergence. Five independent GA runs with different initialisation starting points were performed and only the best individuals kept.

3.2. Software and data analysis

All calculations were performed using MATLAB version 2007b (The MathWorks Inc., Natick, MA); the LibSVM toolbox was used for SVM classifications. The parallel GA scripts developed in-house are based on the Distributed Computing Toolbox (MathWorks) for the parallelisation, the Genetic and Evolutionary Algorithm Toolbox (developed by Hartmut Pohlheim, GEATbx, Berlin) for the basic evolutionary operators, the PLS Toolbox (Eigenvector Research, Manson, WA) for pre-processing and the LibSVM toolbox (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm) for the SVM classification. The programme runs on a High Performance Computing Cluster (Transtec AG, Tübingen) with one master computer and 4 computing nodes using 2 dual core AMD Opteron processors. With the parallel GA running on this 16 processor architecture, a speed increase by a factor of around 12 is generally observed as compared to that achieved using a single processor.

4. Results and discussion

4.1. NIR spectral dataset

The optimised parameters using GA and the error of classification given by the SVM model on the NIR dataset are shown in

Table 1
Classification results on the NIR dataset.

Model name	Selected pre-processing	CV (%)	Test (%)	Sensitivity	Selectivity
PLS-DA (8LV)	1Der, 13 point gap		82.8	92.8	81.5
SVM-mnc	Mean centre	84.3	85.1	84.3	85.0
GA1	Detrend order 3	85.2	86.3	82.8	86.7
GA2	SNV + 1Der(2,21)	86.5	87.8	87.1	88.0
GA3	1Der(2,21) + normalise + detrend order 3	86.8	87.4	87.1	87.4
GA4	Weighted least square order one + 1Der(2,21) + SNV + normalise	86.8	87.8	85.7	88.3

LV, latent variables; CV, cross-validation; Test, test set; 1der, first derivative.

Table 2
Comparison of method efficiency on NIR dataset with McNemar's test.

Method/Test (%)	PLS-DA 82.8	SVM-mnc 85.1	GA1 86.3	GA2 87.8	GA3 87.4	GA4 87.8
PLS-DA	N.A.	4.51	5.78	13.10	10.8	12.5
SVM-mnc		N.A.	0.44	3.01	2.10	2.74
GA1			N.A.	1.02	0.57	0.70
GA2				N.A.	0.04	0
GA3					N.A.	0

Table 3
Classification results on the FTIR spectral dataset.

Model name	Selected pre-processing	CV%	Test%	Sensitivity	Selectivity
PLS-DA 15LV	1Der, 13 point gap		78.2	79	78
SVM-mnc	Mean centre	73.5	74.7	77.1	74.4
GA1	1Der(3, 21)	77.0	78.6	77.2	78.7
GA2	MSC + smoothing (3,15)	79.3	81.1	81.4	81.1
GA3	Smoothing (2,15) + 1Der(3,21) + normalise	80.7	82.2	84.3	82.0
GA4	SNV + smoothing (2,7) + smoothing (3,7) + 1Der(3,21)	81.0	82.7	85.7	82.3

LV, latent variables; CV, cross-validation; Test, test set; 1Der, first derivative, smoothing: Savitzky–Golay smoothing.

Table 4
McNemar's test results for the FTIR dataset.

Method/Test (%)	PLS-DA 78.2	SVM-mnc 74.7	GA1 78.6	GA2 81.1	GA3 82.2	GA4 82.7
PLS-DA	N.A.	3.52	0.15	2.65	4.23	5.39
SVM-mnc		N.A.	3.74	10.56	14.20	15.59
GA1			N.A.	1.84	3.33	4.34
GA2				N.A.	0.44	0.98
GA3					N.A.	0.12

Table 1. The maximum number of consecutive pre-processing steps was fixed between 1 and 4 (called GA1–GA4). Results of the best PLS-DA model (previously published by Woodcock et al., 2008) and the results obtained with SVM on the mean centred spectra (SVM-mnc) are given for comparison. Effectively when using SVM on spectroscopic data the pre-processing is generally not optimised and mean centre is often used as the default one.

The use of a non-linear classification method (in this case SVM with RBF kernel) gives better predictive ability on the test set than the linear PLS-DA method and increases slightly with the number of pre-processing steps. Furthermore, it can be seen that sensitivity and selectivity values are quite similar for all SVM models which is not the case for PLS-DA (10% more for the Ligurian class).

When a single pre-processing method was selected by the GENOPT-SVM method (GA1), this was a procedure correcting multiplicative effects caused by scatter (i.e. SNV or detrend) (Barnes, Dhanoa, & Lister, 1989); in the case of the GA2 method, a first derivative step was used in addition to the SNV. A normalisation procedure was additionally selected when 3 or 4 successive pre-processing steps were applied.

When comparing the correct classification accuracy of the best PLS-DA model and the SVM models (Table 2), it may be observed that all SVM models give a significantly higher accuracy than the best PLS-DA model (McNemar value's greater than 3.841). Furthermore the predictive ability of the SVM model with mean centre spectra and the SVM models with optimised pre-processing were not significantly different. Therefore it seems that for this dataset the pre-processing had only a slight effect on classification efficiency. Finally one pre-processing step is enough to produce SVM models with good predictive ability.

4.2. FTIR spectral dataset

Results of the GENOPT-SVM classification for different values of the maximum number of consecutive pre-processing steps (up to 4), the SVM model on mean centre spectra and the best PLS-DA model obtained by Hennessy et al. (2009) are shown in Table 3. These different classification models were also compared using McNemar's test (Table 4).

The SVM model built using mean centred spectra gave smaller overall accuracy than the one for PLS-DA with optimised pre-processing (74.7% compared to 78.2% for the test set). This accuracy is slightly higher than the one obtained for the same pre-processing (mean centre) using PLS-DA (72%) (Hennessy et al., 2009).

Nevertheless when the pre-processing is optimised for the SVM model, higher overall accuracy than PLS-DA was observed. When only one or two pre-processing steps are used before SVM classification, the prediction ability of these models was not significantly different than that for PLS-DA model (Table 4). When more than 2 pre-processing steps are used, however, the difference became significant. For all these SVM models, which combine Savitzky–Golay smoothing and first derivative pre-processing, the percentage of correct classification on the test set is greater than 82% while the balance between sensitivity and selectivity are comparable to the ones reported for the PLS-DA model. This example demonstrates that pre-processing optimisation is necessary to reveal the classification potential of SVM models. Moreover, in order to assess the importance of the pre-processing order, new SVM models using the 3 pre-processing method selected by GA3 but in a different order have been used. The results are presented Table S3. The higher correct classification accuracies (CV% and Test%) are those obtained with the pre-processing order selected by the GENOPT-SVM procedure. Therefore, the order of the pre-processing steps is important and the order selected by the GENOPT-SVM was the one leading to the best model in terms of predictive ability.

5. Conclusions

In this article a method, based on genetic algorithm, for simultaneous optimisation of data pre-processing and SVM meta–parameters has been proposed. To test the performance of this method the same classification task using NIR or FTIR spectra has been used and the results of the optimisation discussed.

In the case of the NIR spectral dataset, a statistically, and probably practically, significant improvement in correct classification accuracy (from 82.8% to 87.8%) over that previously reported using PLS-DA was achieved by the use of the GENOPT-SVM procedure reported in this paper. It is noteworthy that most of this accuracy improvement (up to 86.3%) occurred using only a single pre-processing step, in this case a detrend procedure. Here the overall accuracy gain, compared to PLS-DA, was mainly due to the use of a non-linear classification method. Only a small improvement was observed when optimised pre-processing was used.

This overall accuracy gain was however at the expense of model sensitivity, with the PLS-DA approach giving a higher value for this parameter (92.8%) than the best GENOPT-SVM procedures (87.1% for GA2 or GA3). Corresponding specificity results for the GENOPT-SVM models (86.7% to 88.3%) were all higher than the corresponding value for the PLS-DA model (81.5%). Therefore, for this dataset, GENOPT-SVM models produced models which were generally better balanced with respect to sensitivity and specificity; this may not, of course be a particular advantage depending on the question being addressed by the analyst. If the goal is to confirm that the NIR spectral signature of a Ligurian olive oil is consistent with a library of authentic Ligurian oil samples, then the model producing the highest sensitivity is to be favoured. This can be easily done with the proposed method by changing the fitness function from overall accuracy to a measure of the sensitivity in order to select relevant SVM models.

For the FTIR spectral dataset, all the classification models produced lower accuracy than those discussed above. For SVM model using mean centred spectra, the accuracy was lower than that obtained for PLS-DA with optimised pre-processed data (74.7% compared to 78.6%). However, when the GENOPT-SVM was used with more than 2 optimised pre-processing steps, significant improvement (82.2% and 82.7% for three and four pre-processes, respectively) was obtained. Therefore for this dataset the optimisation of the pre-processing was crucial for the development of SVM models. Furthermore it has been shown than the pre-processing order selected by GENOPT-SVM was the one leading to the best model. It is interesting to note that accuracy gains were made in both sensitivity and specificity which remained in close agreement for all models. In this case, therefore, the GA3 model represents a useful increase in overall accuracy, specificity and sensitivity over the PLS-DA method.

On the two datasets, it has been shown that both SVM and pre-processing optimisation was required to obtain SVM models with good overall accuracy. Therefore their simultaneous optimisation is needed and this can be done with GENOPT-SVM.

Furthermore the proposed method can be easily extended. For example if the use of a dimension reduction method is required a new step can be added and the number of principal components can be optimised using the genetic algorithm approach.

Acknowledgements

Spectral data used in this study was provided by Gerard Downey and originated in the project "TRACE: Tracing food commodities in Europe" funded under by the EU Sixth Framework Programme (EU IP 006942).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.foodchem.2013.10.020.

References

Balabin, R. M., Safieva, R. Z., & Lomakina, E. I. (2010). Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. *Analytica Chimica Acta, 671*(1–2), 27–35.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy, 43*(5), 772–777.

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2006). *Nonlinear programming: Theory and algorithms*. Wiley-Interscience.

Berrueta, L. A., Alonso-Salces, R. M., & Heberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A, 1158*(1–2), 196–214.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning, 46*(1–3), 131–159.

Cristianni, N., & Shawe-Taylor, J. (2000). *Support vector machines and other kernel-based learning methods*. Cambridge University Press.

Devos, O., & Duponchel, L. (2011). Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression. *Chemometrics and Intelligent Laboratory Systems, 107*(1), 50–58.

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L., & Huvenne, J. P. (2009). Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and intelligent Laboratory systems, 96*(1), 27–33.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley Professional.

Hennessy, S., Downey, G., & O'Donnell, C. P. (2009). Confirmation of food origin claims by Fourier transform infrared spectroscopy and chemometrics: Extra virgin olive oil from Liguria. *Journal of Agricultural and Food Chemistry, 57*(5), 1735–1741.

Holland, J. H. (1992). *Adaptation in natural and artificial systems*. A Bradford Book.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157.

Mitchell, M. (1998). *An introduction to genetic algorithms*. A Bradford Book.

Momma, M. & Bennett, K. P. (2002). A pattern search method for model selection of support vector regression. *2nd SIAM International Conference on Data Mining* (*SDM 02*), Arlington, Va, Siam.

Rinnan, A., Berg, F. V. D., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry, 28*(10), 1201–1222.

Roggo, Y., Duponchel, L., Ruckebusch, C., & Huvenne, J. P. (2003). Statistical tests for comparison of quantitative and qualitative models developed with near infrared spectral data. *Journal of Molecular Structure, 654*(1–3), 253–262.

Schölpkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, Massachusetts/London, England: MIT press.

Talbi, E. G. (2009). *Metaheuristics: From design to implementation*. Wiley.

Tseng, P., & Yun, S. W. (2010). A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Computational Optimization and Applications, 47*(2), 179–206.

Woodcock, T., Downey, G., & O'Donnell, C. P. (2008). Confirmation of by NIR declared provenance of European extra virgin olive oil samples spectroscopy. *Journal of Agricultural and Food Chemistry, 56*(23), 11520–11525.

Xu, Y., Zomer, S., & Brereton, R. G. (2006). Support vector machines: A recent method for classification in chemometrics. *Critical Reviews in Analytical Chemistry, 36*(3–4), 177–188.