Pattern Recognition Letters
journal homepage: www.elsevier.com

A spectral envelope approach towards effective SVM-RFE on infrared data

Flavio E. Spetale[a,b,**], Pilar Bulacio[a,b], Serge Guillaume[c], Javier Murillo[a,b], Elizabeth Tapia[a,b]

[a] *CIFASIS-Conicet, 27 de Febrero 210 bis, S2000EZP Rosario, Argentina*
[b] *Facultad de Ciencias Exactas, Ingeniera y Agrimensura, Universidad Nacional de Rosario, S2000EZP Riobamba 245 bis, Rosario, Argentina*
[c] *Irstea, 361 rue Jean-Franois Breton, F-34196 Montpellier Cedex 5, France*

ABSTRACT

Infrared spectroscopy data is characterized by the presence of a huge number of variables. Applications of infrared spectroscopy in the mid-infrared (MIR) and near-infrared (NIR) bands are of widespread use in many fields. To effectively handle this type of data, suitable dimensionality reduction methods are required. In this paper, a dimensionality reduction method designed to enable effective Support Vector Machine Recursive Feature Elimination (SVM-RFE) on NIR/MIR datasets is presented. The method exploits the information content at peaks of the spectral envelope functions which characterize NIR/MIR spectra datasets. Experimental evaluation across different NIR/MIR application domains shows that the proposed method is useful for the induction of compact and accurate SVM classifiers for qualitative NIR/MIR applications involving stringent interpretability or time processing requirements.

1. Introduction

Infrared (IR) spectroscopy is a non-invasive technique allowing the identification and characterization of chemical compounds using their interaction with light. Applications of IR spectroscopy in the mid-infrared (MIR) and near-infrared (NIR) bands are of widespread use in many fields, including agriculture (Ge et al., 2011; Rossel et al., 2006), food and wines quality (Ferreira et al., 2015; Fudge et al., 2011; Li et al., 2007), postharvest handling of fruits and vegetables (Beckles, 2012; Nicola et al., 2007) and plastic recycling (Kassouf et al., 2014).

Main advantages and limitations of MIR and NIR techniques can be explained by the differences in the origin of their absorption spectra. While the MIR spectra follow from the vibration of fundamental bands, the NIR spectra follow from the overtone and combination of fundamental MIR bands. Hence, while the MIR spectra tend to be simple with very sharp and specific peaks, the NIR spectra tend to be rather complex with many broad overlapping bands. Thus, the interpretation of NIR spectra can be very challenging, especially for complex mixtures of samples. However, since the absorption of light in the NIR region (780-2500 nm) is less intense than in the MIR one (2500-15000 nm), a deeper penetration of light into matter can be accomplished and a minimal sample preparation is required for NIR applications.

In practice, IR spectra are presented as high dimensional vectors of factors. For the NIR case, factors are highly correlated. To effectively handle this type of data, dimensionality reduction methods are required. For quantitative applications, with main focus on predictive modeling and not on the identification of relations between factors, Partial Least Squares (PLS) regression methods (Mehmood et al., 2012) are traditionally used. Briefly, by means of PLS regression methods, a handy number of latent factors accounting for most of the variation of target responses are first selected and then used to perform linear predictions. On the other hand, for qualitative applications, with main focus just on the identification of robust classification boundaries (Langeron et al., 2007), PLS-DA (Boulesteix and Strimmer, 2007; Gurdeniz and Ozen, 2009) methods can be applied. However, when interpretability is also required feature selection methods, allowing the identification of relevant classification factors, must be used (Suphamitmongkol et al., 2013). This is especially true for almost real-time qualitative NIR applications based on Support Vector Machines (SVM) classifiers (Boser et al., 1992), a class of machine learning algorithms characterized by their high accuracy and its ability for modeling diverse types of high dimensional data (Vapnik, 2005). Applications of SVMs can be found in multiple fields, including bioinformatics (Ramaswamy et al., 2001), sound analysis

** Corresponding author: Tel.: +54-341-423-7248; fax: +54-341-482-1772;
e-mail: spetale@cifasis-conicet.gov.ar (Flavio E. Spetale)

(Guo and Li, 2003) and chemometrics (Xu et al., 2006). Owing to the natural ability of SVMs classifiers to deal with high dimensional data, initial works with SVMs in chemometrics focused more on model selection than on data interpretation or time-processing issues (Chen et al., 2007; Devos et al., 2009), i.e., the complete spectrum of IR datasets were usually considered. However, to accomplish compact and thus interpretable SVM classifiers for almost real-time qualitative applications, a reduced fraction of the IR spectra is required. From the application point of view, working with specific regions instead of the complete spectrum would allow the utilization of IR sensors of higher resolution. To this aim, we first note that the highly correlated nature of the NIR spectra limits the effectiveness of fast univariate feature selection methods assuming the independence between features (Saeys et al., 2007). Actually, to avoid the selection of redundant features that may be induced by univariate methods, multivariate feature selection, able to take into account interaction between features are recommended. We note, however, multivariate feature selection methods dismiss specific learning aspects of classification methods, a critical aspect in the construction of compact and accurate SVM classifiers.

To introduce specific learning aspects of classification methods into feature selection tasks, embedded feature selection methods are required. For SVM classifiers this can be accomplished with the SVM recursive feature elimination (SVM-RFE) (Guyon et al., 2002) method, a feature selection method built upon SVM classifiers aiming to identify relevant feature subsets. We note, however, that few studies have considered the direct application of SVM-RFE to the problem of NIR samples classification. As mentioned in (Deng et al., 2013), SVM-RFE can be too computationally expensive, specially when only one least useful feature is removed at each iteration step. Also, SVM-RFE may be unstable with respect to variations in the training data (Kalousis et al., 2007). Although of both these problems may be mitigated with SVM-RFE ensemble variants (Tapia et al., 2012), we note that SVM-RFE does not specifically consider the redundancy between features (Mundra and Rajapakse, 2010). Hence, SVM-RFE on IR datasets may lead to the selection of redundant wavelengths and this undesirable effect may be just reinforced by SVM-RFE ensemble variants. Since direct application of SVM-RFE to IR datasets may be suboptimal, alternative feature selection methods based on genetic algorithms (Ghasemi-Varnamkhasti and Forina, 2014; Moscetti et al., 2014) and random forest classifiers with PCA (Yuhua et al., 2013) have been reported in literature. These considerations strongly suggest that further processing to IR datasets is required before effective SVM-RFE can be accomplished.

In this paper, we show that preservation of the so-called spectral envelope function, a smooth (slowly varying) function of frequency which passes through most significant spectral peaks of IR training datasets, plays an important role in the design of compact and accurate SVM classifiers for qualitative IR applications. With this aim, a two-stage feature selection algorithm designed to capture main features of the spectral envelope function is presented. For this propose, a set of prospective, yet raw, spectral regions is first identified using an unsupervised approach around most significant IR peaks of the spectral envelope function. These regions are further refined using an stabilized version of the SVM-RFE algorithm with respect to variations in the training data. To favor interpretability issues, spectral regions are *individually* refined. In this way, core spectral envelope information gets preserved. The complete set of spectral points across refined IR regions is then used to train compact SVM classifiers.

2. Spectral envelope functions towards effective SVM-RFE on IR data

We notice that the problem of selecting a reduced set of discriminative wavelengths for challenging qualitative NIR applications closely resembles that of the fundamental frequency estimation of a mixture of harmonic sources in the context of music applications (Poliner et al., 2007; Casey et al., 2008). We observe that in the audio setting, data is often reduced for retaining salient information while omitting peripheral details. A strong data reduction technique of music signals is the representation of the full signal spectra to observed spectral peaks (Duan et al., 2010). The usefulness of this approach stems from at least two facts: it is largely known that resynthesis of harmonic sounds from observed spectral peaks cause little changes in human perception (Smith and Serra, 1987) and for harmonic sounds, spectral peaks tend to appear at integer multiples of target fundamental frequencies. Spectral peaks define the spectral envelope. As pointed out by Duan et al. (2008), significant peaks are required to be higher than a baseline, a kind of noise floor so that peaks under such baseline have high probabilities of being generated by noise. On the other hand, it is widely known that for quantitative IR applications, peaks of the IR spectrum are associated with characteristic vibrations of specific functional groups and thus, their heights are proportional to concentration of chemical species in samples (Smith, 1998; Stuart, 2005). Under these considerations, it follows that for qualitative IR applications, IR datasets may be characterized by spectral envelope functions and that these functions may be valuable for extracting potentially discriminative wavelengths, i.e., wavelengths associated with harmonics of core fundamental frequencies.

2.1. Unsupervised learning of IR spectral envelope functions

Let us consider a IR dataset D containing m training samples, each sample characterized by n wavelengths, i.e., $D = \left\{d_i^j,\ i = 1 \ldots m, j = 1 \ldots n\right\}$. The raw spectral envelope function E induced by D (see Fig. 1-a) is given by Eq.1:

$$E\left(x_j\right) = y_j = \max_{i \in 1 \ldots m} d_i^j \quad j \in 1 \ldots n \tag{1}$$

The raw spectral envelope function E is then processed for the unsupervised identification of significant peaks. Hence, all wavelengths below a baseline $b = median\left(\left\{y_j, j = 1 \ldots n\right\}\right)$ are set to b (see Fig. 1-b); the choice of median rather than mean of E aims to overcome the well-known sensitivity of the mean to outliers. As a result, a truncated spectral envelope function E^* is obtained:

$$E^*(x_j)\begin{cases} y_j & y_j > b \quad \forall j \in 1 \ldots n \\ b & \text{otherwise} \end{cases}$$

The truncated spectral envelope function E^* is then inspected for the identification of the set P of wavelengths x_p associated with local maximums of E^*. In addition, the set M of wavelengths associated with local minimums of E^* is also computed.

2.2. Unsupervised identification of spectral windows

Taking into account the nature of the IR spectra, we expect that broad peaks of the truncated spectral envelope function E^* contains important harmonics of core fundamental frequencies. Aiming to accomplish a compact representation of the IR spectra, the truncated spectral envelope function E^* is used to guide the identification of significant spectral regions, hereafter called spectral windows. For this purpose, the Windows from Envelope (WE) algorithm (see Algorithm 1) is introduced.

Given a training IR dataset D, WE first computes the raw spectral envelope function E *(L.4)*, continues with a baseline b *(L.6)* and then its truncated version E^* with baseline b *(L.8)*. From E^*, the corresponding sets P of local maximus *(L.13)* and the set M of local minimums *(L.14)* are computed. For each $x_p \in P$, WE identifies the spectral window *(L.16)* centered on x_p with width $w_p = \left(x_p^r - x_p^l\right)$ (see Fig. 1-c), where x_p^r and x_p^l are respectively the right and left closer wavelengths to x_p where E^* falls to $Max\left[b, decay * E^*\left(x_p\right)\right]$. The *decay* parameter, $0 < decay \leq 1$, is used to control spectral window widths. For sharp E^* peaks, very narrow spectral windows are obtained despite the specific setting of the *decay* parameter. The resulting set of spectral windows is further processed for additional dimensionality reduction using the information about local minimums of E^* available in M. Hence, narrower windows w_p^* *(L.17)* are obtained by performing descendant walks from wavelengths x_p until the first local minimum of E^*, if any, is found, $p = 1 \ldots |P|$ (see Fig. 1-d). Afterwards, the final set of spectral windows F *(L.19)* is built from P and W^*.

2.3. Supervised SVM-RFE refinement of spectral windows

SVM-RFE makes feature selection using a backward elimination process based on the weights computed by a linear SVM classifier. To deal with small variations in the training data, a robust version, built upon a 5-Fold CV approach and called SVM-RFE*, was proposed by the authors Tapia et al. (2012). To further refine the training dataset obtained after WE processing, SVM-RFE* was applied *to each* spectral window. The rationale behind this decision is twofold. The first reason is relate to computationally load, i.e., SVM-RFE scales quadratically with the number of features and thus, its application on a per-window basis reduces computational complexity by a factor proportional to the number of spectral windows. The second reason is related to the importance of spectral envelope functions in the characterization of IR datasets. Note that applying SVM-RFE to whole and fused set of spectral windows may drop key wavelengths for the definition of the spectral envelope function. Hence, if this function is indeed important for the characterization of IR datasets, its main features must be preserved. This objective can only be accomplished if SVM-RFE* is applied in a per-window basis mode.

Based on the above considerations, spectral windows were individually refined with an additional SVM-RFE* processing stage using a 5-Fold CV setup. Therefore, for each cross-validation run and for each SVM-RFE iteration step, a validation error was obtained using four folds for training and one fold for validation. At the end of SVM-RFE iterations, the mean validation error was computed and the smallest feature subset with a validation error below such mean was selected. Aiming to promote feature selection stability, only those features selected in the 5 cross-validation runs were finally selected. The union of feature subsets obtained for each spectral window was then used to build a reduced training dataset.

2.4. Sensitivity analysis of SVM-RFE refinement

In order to set the *decay* parameter, we analyzed its sensitivity to the combination of the WE algorithm and robust SVM-RFE (WE+SVM-RFE*) with respect to variations in the training data. To this aim, the fraction of preserved features along with their stability and the classification accuracy of corresponding linear SVM classifiers were evaluated for different settings of the *decay* parameter in the range $[0.5, 0.9]$. Regarding the stability of feature selection, the similarity index I_s proposed by (Kalousis et al., 2005) was used. Given two subsets of features A and B, respectively obtained with *decay* parameters d_A and d_B, the similarity between both subsets is given by $I_s = \frac{|A \cap B|}{|A \cup B|}$. To perform evaluations, a 5-Fold CV approach on the two following IR datasets was considered:

Diesel: This dataset, obtained from data in Kalivas (1997), contains 60 NIR samples of three types of gasoline (17, 23 and 20 samples) defined by their octane number. Each NIR sample consists of 401 wavelengths in the range of 900-1700 nm.

Wine: This dataset, provided by Marc Meurens[1], contains 124 MIR samples of three types of wine (37, 36 and 48 samples) defined by their alcohol level. Each MIR sample consists of 252 wavelengths in the range of 400-4000 cm^{-1}.

Average 5-Fold CV results on the two datasets for the fraction of selected features (see Fig. 2a) and the classification accuracy of corresponding SVM classifiers (see Fig. 2b) suggested that a *decay* parameter between 0.65 and 0.8 may lead to satisfactory performance results. To make a final decision on a robust value for the *decay* parameter, feature selection stability results (see Tables 1 and 2) were analyzed. Hence, we searched for *decay* pairs in the grid $[0.5, 0.9] \times [0.5, 0.9]$ showing the highest I_s values with the smallest variations near the diagonals. As a result of this analysis, even if other values are also possible, the *decay* parameter was set to 0.75.

[1] http://mlg.info.ucl.ac.be/index.php?page=DataBases

Algorithm 1 Windows from Envelope (WE) algorithm

INPUT: IR dataset D with m training samples and n wavelengths, $D = d_i^j$ with $i = 1 \dots m$, $j = 1 \dots n$, parameter *decay*.
OUTPUT: A set of spectral windows F.
for $j \in n$ **do**
    $E(x_j) =$ ***max*** (d_i^j) // Compute the envelope E from D
end for
$b =$ ***median*** $(E(x_j))$ // Compute the baseline b from E
for $j \in n$ **do**
    if $E(x_j) > b$ **then**
        $E^*(x_j) = E(x_j)$ // Compute the truncated envelope E^* with the median b of E
    else $E^*(x_j) = b$
    end if
end for
$P \leftarrow$ ***maximums*** (E^*) // Compute the set P of local maximums of E^*
$M \leftarrow$ ***minimums*** (E^*) // Compute the set M of local minimums of E^*
for $p \in P$ **do**
    $w_p \leftarrow$ ***widths*** $(P, E^*, decay)$ // Compute window widths w_p centered on $x_p \in P$
    $w_p^* \leftarrow$ ***narrow-widths*** (w_p, E^*, M) // Compute the set W^* of final window widths w_p^* using M
end for
$F \leftarrow$ ***build-windows*** (P, W^*) // Compute the final set F of spectral windows from P and W^*

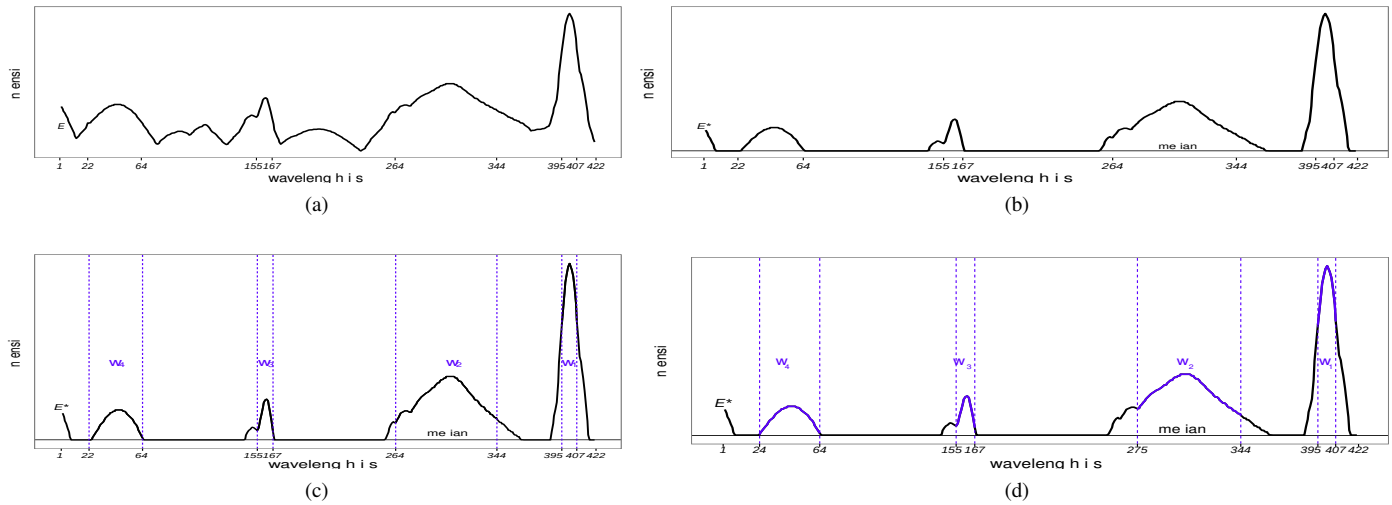


Fig. 1: The unsupervised spectral envelope approach for IR data dimensionality reduction. (a) The raw spectral envelope function E is induced from local maximums of IR dataset. (b) The truncated spectral envelope function E^* is obtained with a baseline $b = median\left(\{y_j, j = 1 \dots n\}\right)$. (c) A set of spectral windows is induced from E^*. (d) Final spectral window widths are computed using the minimums of E^*.

Table 1: WE+SVM-RFE* feature selection stability on the *Diesel* dataset for different settings of the *decay* parameter. Average 5-Fold CV values of the Kalousis index I_s are reported for *decay* parameter pairs (d_a, d_b) on the grid $[0.5, 0.55, \dots, 0.9] \times [0.5, 0.55, \dots, 0.9]$.

$d_a \backslash d_b$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.50	1.00	0.45	0.44	0.41	0.42	0.44	0.36	0.37	0.31
0.55	0.45	1.00	0.49	0.45	0.42	0.41	0.34	0.34	0.32
0.60	0.44	0.49	1.00	0.77	0.69	0.63	0.52	0.56	0.42
0.65	0.41	0.45	0.77	1.00	0.78	0.69	0.58	0.57	0.48
0.70	0.42	0.42	0.69	0.78	1.00	0.82	0.65	0.63	0.52
0.75	0.44	0.41	0.63	0.69	0.82	1.00	0.79	0.63	0.50
0.80	0.36	0.34	0.52	0.58	0.65	0.79	1.00	0.65	0.47
0.85	0.37	0.34	0.56	0.57	0.63	0.63	0.65	1.00	0.62
0.90	0.31	0.32	0.42	0.48	0.52	0.50	0.47	0.62	1.00

3. Numerical experiments

3.1. Description of used datasets

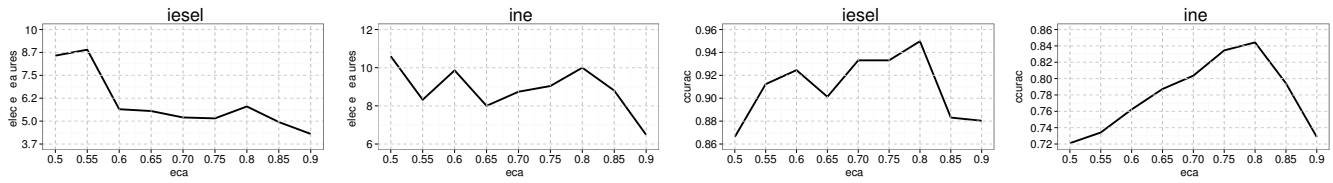
Multiple datasets across different IR domains were selected for evaluating the performance of the WE+SVM-RFE* feature selection algorithm in the construction of accurate and interpretable linear SVM classifiers.

Polymers: This dataset was given from the XXX Project with XXX[2] contains NIR samples of four types of plastic bottles, namely, PET (47 samples), PEHD (125 samples), Polypropylene (50 samples) and PVC (89 samples). In order to be self-

[2] http://www.ondalys.fr/

Table 2: WE+SVM-RFE* feature selection stability on the *Wine* dataset for different settings of the *decay* parameter. Average 5-Fold CV values of the Kalousis index I_s are reported for *decay* parameter pairs (d_a, d_b) on the grid $[0.5, 0.55, \ldots, 0.9] \times [0.5, 0.55, \ldots, 0.9]$.

$d_a \setminus d_b$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.50	1.00	0.49	0.49	0.43	0.39	0.35	0.26	0.19	0.14
0.55	0.49	1.00	0.41	0.41	0.36	0.31	0.22	0.18	0.11
0.60	0.49	0.41	1.00	0.67	0.41	0.51	0.40	0.23	0.12
0.65	0.43	0.41	0.67	1.00	0.50	0.68	0.40	0.27	0.11
0.70	0.39	0.36	0.41	0.50	1.00	0.63	0.41	0.35	0.14
0.75	0.35	0.31	0.51	0.68	0.63	1.00	0.65	0.37	0.12
0.80	0.26	0.22	0.40	0.40	0.41	0.65	1.00	0.51	0.21
0.85	0.19	0.18	0.23	0.27	0.35	0.37	0.51	1.00	0.37
0.90	0.14	0.11	0.12	0.11	0.14	0.12	0.21	0.37	1.00



(a) (b)

Fig. 2: (a) The fraction of selected features by the WE+SVM-RFE* method on the *Diesel* and *Wine* datasets for different settings of the *decay* parameter. Average 5-Fold CV values are reported for *decay* in the range $[0.5, 0.55, \ldots, 0.9]$. (b) SVM classification accuracy on the *Diesel* and *Wine* datasets. Average 5-Fold CV precision values are reported when WE+SVM-RFE* feature selection is performed with the *decay* parameter in the range $[0.5, 0.55, \ldots, 0.9]$.

contained, a brief description of sample collection is made. NIR samples were obtained using a reflexion setup with a halogen light source set to irradiate plastic bottles and a white screen behind them to reflect the light. NIR spectra were acquired using a StellarNet spectrometer (950 to 1700 nm, Black comet model, 256 pixels) controlled by a computer via USB. The wavelength region was chosen because it contains several plastic absorption bands. Bottles were placed with the head below on a moving metallic stick and measurements were performed on the bottom of the bottle in order to reduce interference and be sure no liquid remained, which would dramatically affect spectral signatures. NIR measurements were performed at 2 nm intervals thus giving 422 wavelengths per sample.

Apricots. This dataset, derived from Bureau et al. (2009), contains 731 MIR samples of apricots of three types (230, 244 and 257 samples) defined by their Brix degree, i.e., by their water-soluble sugar concentration. MIR samples consist of 292 wavelengths in the range of 900-1500 cm^{-1}.

Strawberry. This dataset[3] contains 983 MIR samples of two types of fruit pures, namely "Strawberry" (632 samples) and "Non-Strawberry" (251 samples) (Holland and Wilson, 1998). In the former case, pures are prepared from fresh whole fruits by the researchers. In the latter case, purées are prepared from diverse collection of other purées, including strawberry adulterated with other fruits and sugar solutions, raspberry, apple, blackcurrant, blackberry, plum, cherry, apricot, grape juice and mixtures of these. MIR samples consisting of 235 wavelengths in the range 899-1802 cm^{-1} were acquired from each purée sample using attenuated total reflectance sampling.

[3] `http://www.ifr.ac.uk/Bioinformatics/MIRFruitPurees.zip`

3.2. Experimental protocol

The effectiveness of the WE+SVM-RFE* feature selection method in the construction of accurate and compact SVM classifiers for IR data was compared against direct SVM-RFE* and four other popular feature selection algorithms mentioned in the literature. Specifically, we first considered the SVM-RFE* approach that eliminates the least useful feature at each iteration step. We also considered Relief (Kira and Rendell, 1992), a well-known feature *subset* selection algorithm known to handle strong dependencies between features and noise, and three entropy-based feature selection algorithms (Mitchell, 1997): Information Gain (InfoGain), Information Gain Ratio (GainR) and Symmetrical Uncertainty (SymmU), all of them assuming independence between features. One of the three methods of entropy-based feature selection, InfoGain, was evaluated.

For the sake of completeness, dimensionality reduction methods were also considered. These methods involve a space transformation which makes hard the interpretation using the initial, raw, features. Nevertheless, they are widely used as they do not require a feature selection process and they are able to exploit the whole information of the input spectra. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Partial Least Squares Discriminant Analysis (PLS-DA) methods share the way the new space is defined: the axes are linear combinations of the raw features. They differ in the way these axes are designed. PCA (Jolliffe, 2002) maximizes the explained variance of the input spectra, the axes are the eigenvectors of $X^T X$. To make a classification, SVM classifiers are evaluated in the new space. LDA (Brito et al., 2013) maximizes the between group variance, B. The axes are the eigenvectors of $T^{-1}B$, where T stands for the total variance matrix. Finally, PLS-DA (Barker and Rayens, 2003), maximizes the covariance between the input spectra and the target. The

axes are computed using iterative algorithms.

A randomized strategy based on 50×5-Fold CV experiments was used to assess the performance of aforementioned feature selection and dimensionality reduction methods. At each CV fold, an inner 5-Fold CV experiment was performed to estimate the optimal number of features in the SVM-RFE, Relief, Info-Gain feature selection methods and the optimal number of components in the PCA dimensionality reduction technique. Feature selection performance was evaluated by the mean number of features selected across the 50 runs of 5-Fold CV experiments. Similarly, linear SVM classifiers built after feature selection, PCA dimensionality reduction, PLS-DA and LDA classifiers were evaluated with the mean classification accuracy.

In practice, the default implementations of the Relief and entropy-based feature selection methods provided in the R package "FSelector" (Romanski, 2014) were used for supervised feature selection. Similarly, the prcomp implementation of PCA algorithm provided in the R Stats package (R-Core-Team, 2014) was used for unsupervised dimensionality reduction. Finally, the R package "plsgenomics" (Boulesteix et al., 2015) was used for optimized LDA classification and "mixOmics" (Cao et al., 2015) was used for PLS-DA classification.

4. Results and discussion

4.1. The importance of the spectral envelope function

To appreciate the importance of the spectral envelope function in the characterization of IR datasets, three operation modes of SVM-RFE* were evaluated: *i*) per-spectral window after WE processing (WE+SVM-RFE*), *ii*) all spectral windows after WE processing (WE+SVM-RFE$^*_a$), and *iii*) the complete set of wavelengths in the original training data.

A 5-Fold CV approach on *Diesel* and *Wine* datasets was considered. Average 5-Fold CV results for the number of selected features and the classification accuracy (see Table 3) of corresponding linear SVM classifiers suggest that using SVM-RFE* in the per-window operation mode is the best data processing strategy and that there is no advantage in applying SVM-RFE* to all spectral windows over all wavelengths in the original training data.

We wonder whether these results may be due to the preservation of wavelengths of the spectral envelope function accomplished by the SVM-RFE* algorithm when used in the per-spectral window operation mode. To shed some light on this issue, selected features in three SVM-RFE* operation modes were mapped to reference spectral windows obtained after WE processing. It was observed that wavelengths of the spectral envelope functions were practically dismissed by SVM-RFE* when used in the all-wavelengths operation mode and, were only partially preserved in the all-spectral windows operation mode (see Tables 4 and 5). Overall, these results suggest that the preservation of main wavelengths of the spectral envelope function accomplished by the WE+SVM-RFE* algorithm is an important issue for the construction of compact and accurate linear SVM classifiers for IR datasets.

4.2. WE+SVM-RFE performance*

To better understand the difficulty of the three classification problems at hand, a 2D visualization analysis was first performed using PCA. Highly overlapped classes, with no clear linear separation boundaries were observed in all cases. These results suggested the need of dimensionality reduction, e.g., by means of PCA, or feature selection before any classification algorithm could be applied.

Regarding interpretability, Table 6 shows firstly that WE+SVM-RFE* leads to the smallest sets of features compared to other alternatives, including SVM-RFE*. Further screening showed that selected features with WE+SVM-RFE*, as opposed to SVM-RFE*, tend to be always contained in a reduced number of spectral regions associated with more salient peaks, which seem to be related to target classes. For instance, in the Polymer dataset the four spectral regions (A, B, C, D) selected by WE+SVM-RFE* method (see Fig. 3) point out main features of the four plastic bottle spectrum (PVC, PET, PEHD and polypropylene) (Cambridge, 2015). On the other hand, features selected by the raw SVM-RFE* are dispersed across the full spectrum (lines in grey), which makes difficult the interpretation. Altogether, these results suggest the usefulness of the proposed method when both interpretability and classification of IR spectrum are of interest.

Regarding accuracy, Table 7 shows that WE+SVM-RFE* yields similar, or even better, results than the concurrent approaches, including the optimized LDA, PLS-DA and PCA based SVM classifiers. The largest gain is for the Apricot dataset.

5. Conclusions

In this paper, a spectral envelope approach towards effective SVM-RFE on IR datasets has been presented. As it happens with music applications, the spectral envelope function provides a high level and compact representation of IR datasets and thus, subject to suitable processing, it may be used to overcome the difficulties found in the direct application of the SVM-RFE method. These considerations motivate the introduction of the Windows from Envelope algorithm allowing the unsupervised identification of a reduced set of spectral windows supporting the spectral envelope function and thus, the effective application of the SVM-RFE method on IR datasets. Taking into account the well-known sensitivity of SVM classifiers to noise and outliers (Boser et al., 1992) and that a variety of noise sources may affect the quality of IR datasets (Xu et al., 2008), an ensemble approach to SVM-RFE was used.

These insights are captured in the WE+SVM-RFE* proposal for feature selection on IR datasets. Experimental results across three different IR application domains (polymers, agriculture and food) demonstrated that spectral regions achieved with WE+SVM-RFE* can shed light on the relation between spectral regions and target classes. Finally, experimental results across three different IR application domains (polymers, agriculture, and food) suggest the usefulness of the proposed method for the construction of compact, interpretable and accurate SVM classification models for qualitative IR applications.

Table 3: The *number* of *features* selected by WE+SVM-RFE*, WE+SVM-RFE$^*_a$ and SVM-RFE* feature selection methods along with the classification *accuracy* accomplished by corresponding linear SVM classifiers in a 5-Fold CV steup.

Dataset	Type	# Features	WE+SVM-RFE* (Accuracy)	WE+SVM-RFE$^*_a$(Accuracy)	SVM-RFE*(Accuracy)
Diesel	NIR	401	16 (0.94)	13 (0.80)	14 (0.80)
Wine	MIR	252	18 (0.82)	19 (0.78)	18 (0.79)

Table 4: WE+SVM-RFE*, WE+SVM-RFE$^*_g$ and SVM-RFE* feature selection in the *Diesel* dataset. Selected wavelengths (ID numbers) are mapped against reference WE spectral windows specified by their lower and upper wavelength limits.

	WE spectral windows			
Feature selection	[122 – 128]	[144 – 150]	[239 – 255]	Outside
WE+SVM-RFE*	{122 – 125, 128}	{146 – 147, 149 – 150}	{245 – 247, 249 – 252}	
WE+SVM-RFE$^*_g$	{124 – 125, 127}	{145 – 146, 148}	{245 – 251}	
SVM-RFE*			{239, 251}	154-157, 164, 228, 230, 232, 236, 386, 388, 390

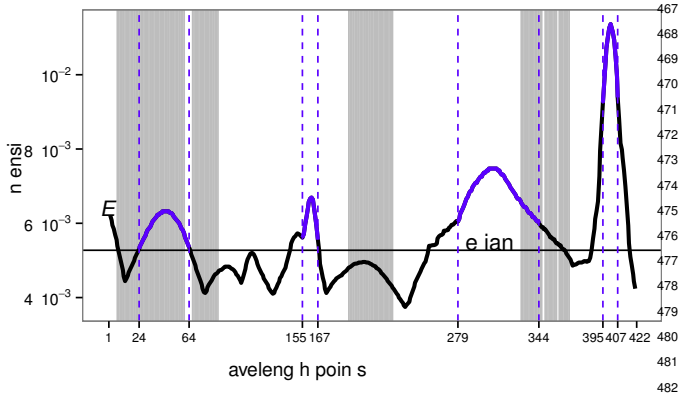


Fig. 3: Polymer envelope function. Dashed blue lines represent the selected regions with WE+SVM-RFE*. Grey lines represent the selected features with SVM-RFE*.

Acknowledgments

The authors were supported by projects PICT PRH No. 0253 (2011) and No. 2513 (2012), ANPCyT, Argentina.

References

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. Journal of Chemometrics 17, 166–173.

Beckles, D.M., 2012. Factors affecting the postharvest soluble solids and sugar content of tomato (solanum lycopersicum l.) fruit. Postharvest Biology and Technology 63, 129 – 140.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, ACM, New York, NY, USA. pp. 144–152.

Boulesteix, A.L., Lacroix, S.L., Peyre, J., Strimmer, K., 2015. plsgenomics. R package version 3.1.0.

Boulesteix, A.L., Strimmer, K., 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 8, 32–44.

Brito, A.L.B., Brito, L.R., Honorato, F.A., Pontes, M.J.C., Pontes, L.F.B.L., 2013. Classification of cereal bars using near infrared spectroscopy and linear discriminant analysis. Food Research International 51, 924–928.

Bureau, S., Ruiz, D., Reich, M., Gouble, B., Bertrand, D., Audergon, J.M., Renard, C.M., 2009. Rapid and non-destructive analysis of apricot fruit quality using ft-near-infrared spectroscopy. Food Chemistry 113, 1323–1328.

Cambridge, U., 2015. Ir spectra for some common polymers. URL: `http://www.doitpoms.ac.uk/tlplib/artefact/polymers.php`. accessed: 2015-10-29.

Cao, K.A.L., Gonzalez, I., Dejean, S., 2015. mixOmics. R package version 3.1.0.

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M., 2008. Content-based music information retrieval: Current directions and future challenges. Proceedings of the IEEE 96, 668–696.

Chen, Q., Zhao, J., Fang, C., Wang, D., 2007. Feasibility study on identification of green, black and oolong teas using near-infrared reflectance spectroscopy based on support vector machine (svm). Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 66, 568–574.

Deng, S., Xu, Y., Li, L., Li, X., He, Y., 2013. A feature-selection algorithm based on support vector machine-multiclass for hyperspectral visible spectral analysis. Journal of Food Engineering 119, 159 – 166.

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, 27 – 33.

Duan, Z., Pardo, B., Zhang, C., 2010. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. Audio, Speech, and Language Processing, IEEE Transactions on 18, 2121–2133.

Duan, Z., Zhang, Y., Zhang, C., Shi, Z., 2008. Unsupervised single-channel music source separation by average harmonic structure modeling. Audio, Speech, and Language Processing, IEEE Transactions on 16, 766–778.

Ferreira, D., Pallone, J., Poppi, R., 2015. Direct analysis of the main chemical constituents in chenopodium quinoa grain using fourier transform near-infrared spectroscopy. Food Control 48, 41–45.

Fudge, A., Wilkinson, K.L., Ristic, R., Cozzolino, D., 2011. Classification of smoke tainted wines using mid-infrared spectroscopy and chemometrics. J Agric Food Chem .

Ge, Y., Thomasson, J., Sui, R., 2011. Remote sensing of soil properties in precision agriculture: A review. Frontiers of Earth Science 5, 229–238.

Ghasemi-Varnamkhasti, M., Forina, M., 2014. {NIR} spectroscopy coupled with multivariate computational tools for qualitative characterization of the aging of beer. Computers and Electronics in Agriculture 100, 34 – 40.

Guo, G., Li, S.Z., 2003. Content-based audio classification and retrieval by support vector machines. IEEE Transactions on Neural Networks 14, 209–215.

Gurdeniz, G., Ozen, B., 2009. Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. Food Chemistry 116, 519–525.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422.

Holland, J. K .and Kemsley, E.K., Wilson, R.H., 1998. Use of fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry pures. Journal of the Science of Food and Agriculture 76, 263–269.

Jolliffe, I.T., 2002. Principal Component Analysis. Second ed., Springer.

Kalivas, J.H., 1997. Two data sets of near infrared spectra. Chemometrics and Intelligent Laboratory Systems 37, 255–259.

Kalousis, A., Prados, J., Hilario, M., 2005. Stability of feature selection algorithms, in: Data Mining, Fifth IEEE International Conference on, p. 8.

Kalousis, A., Prados, J., Hilario, M., 2007. Stability of feature selection algorithms: A study on high-dimensional spaces. Knowl. Inf. Syst. 12, 95–116.

Kassouf, A., Maalouly, J., Rutledge, D.N., Chebib, H., Ducruet, V., 2014. Rapid discrimination of plastic packaging materials using {MIR} spectroscopy coupled with independent components analysis (ica). Waste Management 34, 2131–2138.

Table 5: WE+SVM-RFE*, WE+SVM-RFE$^*_g$ and SVM-RFE* feature selection in the *Wine* dataset. Selected wavelengths (ID numbers) are mapped against reference WE spectral windows specified by their lower and upper wavelength limits

	WE spectral windows							
Feature selection	[24 – 28]	[33 – 37]	[83 – 90]	[93 – 108]	[117 – 126]	[129 – 133]	[202 – 205]	Outside
WE+SVM-RFE*	{25, 27}	{34}	{85, 88 – 90}	{93 – 94, 104, 106, 108}	{117, 120, 126}	{130}	{204 – 205}	
WE+SVM-RFE$^*_g$			{84, 87 – 88}	{96 – 100}	{117 – 120}	{129 – 130, 133}	{202 – 205}	
SVM-RFE*	{24 – 28}							169-170,173...

Table 6: The *number* of *features* selected by the WE+SVM-RFE*, SVM-RFE*, InfoGain, GainR, SymmU and Relief feature selection methods on benchmark NIR/MIR datasets. In the fourth column between brackets the number of selected regions of the spectrum is expressed.

Dataset	Type	# Features	WE+SVM-RFE*(#Regions)	SVM-RFE*	InfoGain	Relief
Polymer	NIR	422	62(4)	90	97	80
Apricot	MIR	292	20(4)	35	78	32
Strawberry	MIR	235	45(6)	70	188	106

Table 7: The classification *accuracy* accomplished by linear SVM classifiers after the application of the WE+SVM-RFE*, SVM-RFE*, InfoGain and Relief feature selection methods and the PCA dimensionality reduction technique to benchmark NIR/MIR datasets. The classification accuracy of optimized LDA and PLS-DA classifiers are shown as reference.

Dataset	T ype	SVM						
		WE+SVM-RFE*	SVM-RFE*	InfoGain	Relief	PCA	LDA	PLS-DA
Polymer	NIR	0.95	0.93	0.95	0.92	0.95	0.93	0.95
Apricot	MIR	0.96	0.85	0.87	0.85	0.90	0.85	0.91
Strawberry	MIR	0.98	0.90	0.98	0.96	0.96	0.96	0.97

Kira, K., Rendell, L.A., 1992. A practical approach to feature selection, in: Proceedings of the Ninth International Workshop on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 249–256.

Langeron, Y., Doussot, M., Hewson, D., Duchłne, J., 2007. Classifying nir spectra of textile products with kernel methods. Engineering Applications of Artificial Intelligence 20, 415 – 427.

Li, X., He, Y., Fang, H., 2007. Non-destructive discrimination of chinese bayberry varieties using vis/nir spectroscopy. Journal of Food Engineering 81, 357–363.

Mehmood, T., Liland, K.H., Snipen, L., Sb, S., 2012. A review of variable selection methods in partial least squares regression. Chemometrics and Intelligent Laboratory Systems 118, 62 – 69.

Mitchell, T.M., 1997. Machine Learning. 1 ed., McGraw-Hill, Inc., New York, NY, USA.

Moscetti, R., Haff, R.P., Saranwong, S., Monarca, D., Cecchini, M., Massantini, R., 2014. Nondestructive detection of insect infested chestnuts based on {NIR} spectroscopy. Postharvest Biology and Technology 87, 88 – 94.

Mundra, P., Rajapakse, J., 2010. Svm-rfe with mrmr filter for gene selection. NanoBioscience, IEEE Transactions on 9, 31–37.

Nicola, B.M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K.I., Lammertyn, J., 2007. Nondestructive measurement of fruit and vegetable quality by means of {NIR} spectroscopy: A review. Postharvest Biology and Technology 46, 99–118.

Poliner, G.E., Ellis, D., Ehmann, A., Gomez, E., Streich, S., Ong, B., 2007. Melody transcription from music audio: Approaches and evaluation. Audio, Speech, and Language Processing, IEEE Transactions on 15, 1247–1256.

R-Core-Team, 2014. The R Stats Package. R package version 3.1.0.

Ramaswamy, S., Tamayo, P., Rifkin, et al, 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences of the United States of America 98, 15149–15154.

Romanski, P., 2014. FSelector. R package version 3.1.0.

Rossel, R.V., Walvoort, D., McBratney, A., Janik, L., Skjemstad, J., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75.

Saeys, Y., Inza, I., Larraaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517.

Smith, B., 1998. Infrared Spectral Interpretation. CRC Press.

Smith, J.O., Serra, X., 1987. PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation, in: International Computer Music Conference (ICMC), International Computer Music Association. pp. 290–297.

Stuart, B.H., 2005. Infrared Spectroscopy: Fundamentals and Applications. John Wiley & Sons, Ltd.

Suphamitmongkol, W., Nie, G., Liu, R., Kasemsumran, S., Shi, Y., 2013. An alternative approach for the classification of orange varieties based on near infrared spectroscopy. Computers and Electronics in Agriculture 91, 87–93.

Tapia, E., Bulacio, P., Angelone, L., 2012. Sparse and stable gene selection with consensus svm-rfe. Pattern Recognition Letters 33, 164–172.

Vapnik, V., 2005. Universal learning technology: Support vector machines. NEC Journal of Advanced Technology 2, 137–144.

Xu, L., Zhou, Y.P., Tang, L.J., Wu, H.L., Jiang, J.H., Shen, G.L., Yu, R.Q., 2008. Ensemble preprocessing of near-infrared (nir) spectra for multivariate calibration. Analytica Chimica Acta 616, 138–143.

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

Yuhua, Q., Xiangqian, D., Huili, G., 2013. Application of high-dimensional feature selection in near-infrared spectroscopy of cigarettes' qualitative evaluation. Spectroscopy Letters 46, 397–402.