Pattern Recognition Letters 71 (2016) 59–65

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

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, Rosario S2000EZP, Santa Fe, Argentina
[b] Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario, S2000EZP Riobamba 245 bis, Rosario, Santa Fe, Argentina
[c] Irstea, 361 rue Jean-François Breton, Montpellier Cedex 5 F-34196, Languedoc-Rosellón, France

ARTICLE INFO

Article history:
Received 29 May 2015
Available online 30 December 2015

Keywords:
Spectral envelope
Infrared spectroscopy
dimensionality reduction

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 [18,41], food and wines quality [16,17,31], postharvest handling of fruits and vegetables [2,36] and plastic recycling [28].

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 [32] 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 [30], PLS-DA [5,21] methods can be applied. However, when interpretability is also required feature selection methods, allowing the identification of relevant classification factors, must be used [46]. This is especially true for almost real-time qualitative NIR applications based on Support Vector Machines (SVM) classifiers [3], a class of machine learning algorithms characterized by their high accuracy and its ability for modeling diverse types of high dimensional data [48]. Applications of SVMs can be found in multiple fields, including bioinformatics [39], sound analysis [20] and chemometrics [50]. 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 [11,13], 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

☆ This paper has been recommended for acceptance by J. Yang

* Corresponding author at: CIFASIS-Conicet, 27 de Febrero 210 bis, S2000EZP Rosario, Argentina. Tel.: +54 341 423 7248; fax: +54 341 482 1772.
E-mail address: spetale@cifasis-conicet.gov.ar (F.E. Spetale).

http://dx.doi.org/10.1016/j.patrec.2015.12.007
0167-8655/

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 [42]. 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) [22] 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 [12], 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 [27]. Although of both these problems may be mitigated with SVM-RFE ensemble variants [47], we note that SVM-RFE does not specifically consider the redundancy between features [35]. 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 [19,34] and random forest classifiers with PCA [51] 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 a 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 [10,37]. 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 [14]. 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 [44] 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. [15], 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 [43,45]. 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., $D = \{d_i^j,\ i = 1 \dots m,\ j = 1 \dots n\}$. The raw spectral envelope function E induced by D (see Fig. 1(a)) is given by Eq. (1)

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

The raw spectral envelope function E is then processed for the unsupervised identification of significant peaks. Hence, all wavelengths below a baseline $b = median(\{y_j, j = 1 \dots n\})$ 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

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

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 maximums *(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 $w_p = (x_p^r - x_p^l)$ (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 $Max[b, decay*E^*(x_p)]$. 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, $p = 1 \dots |P|$ (see Fig. 1(d)). Afterwards, the final set of spectral windows F *(L.19)* is built from P and W^*.