Supervised Variable Clustering for Classification of NIR Spectra

Catherine Krier[1]*, Damien François[2], Fabrice Rossi[3], Michel Verleysen[1],

[1] Université catholique de Louvain, Machine Learning Group,
place du Levant, 3, 1348 Louvain la Neuve, Belgique
{catherine.krier, michel.verleysen}@uclouvain.be
[2] Université catholique de Louvain, Machine Learning Group,
av. G. Lemaître, 4, 1348 Louvain la Neuve, Belgique
Damien.francois@dmf-solutions.be
[3] Institut TELECOM, TELECOM ParisTech, LTCI - UMR CNRS 5141
46, rue Barrault, 75013 Paris, France
Fabrice.Rossi@apiacoa.org

Abstract. Spectrometric data involve very high-dimensional observations representing sampled spectra. The correlation of the resulting spectral variables and their high number are two sources of difficulties in modeling. This paper proposes a supervised feature clustering algorithm that provides dimension reduction for this type of data in a classification context. The new features designed by this method are means of the original spectral variables computed on specific ranges of wavelengths and are therefore easy to interpret. Experiments on real world data show that the reduction in redundancy and in number of features leads to better performances obtained using a very low number of spectral ranges.

1 Introduction

Near-infrared (NIR) spectra are generally given as high-dimensional vectors obtained by high resolution sampling of the underlying smooth spectra. The corresponding spectral variables are correlated but distinct enough to generate difficulties linked to the "curse of dimensionality". Some dimensionality reduction is therefore clearly needed. However many classical solutions have the drawback of producing new features which are difficult to interpret in terms of the original spectral variables. While those algorithms can lead to very good performances, they do not provide any knowledge on e.g. which part of spectrum is relevant for the given task, which is essential for some industrial applications.

This paper addresses these problems in the specific case of spectra classification via a combination of a supervised dimensionality reduction method and a feature selection approach. The first step consists in clustering adjacent spectral variables (as proposed in [1] and [2] for prediction problems). Contrarily to these clustering approaches, the proposed criterion used to group the variables is supervised by the class label. The dimensionality of the data is therefore reduced in an interpretable way as each group of spectral variables is replaced by its mean: this corresponds roughly to a data dependent downsampling of the observed spectra. Then a feature selection algorithm is applied to retain a limited number of clusters of spectral variables: a

* The work of C. Krier is funded by a Belgian FRIA grant.

wrapper approach is used to build a Support Vector Machine (SVM) on a subset of the clusters selected in the previous step.

The rest of the paper is organized as follows. Section 2 describes the clustering-based methodology proposed in this paper and Section 3 illustrates this methodology on two real world datasets.

2 Methodology

The proposed methodology consists in two major elements: a feature clustering method followed by a feature selection method. In this paper, we proceed as follows:

1. a clustering hierarchy is built on the spectral variables
2. an optimal level in the hierarchy is chosen by maximizing a cross-validation based estimation of the performances of a linear discrimination model constructed on the reduced representation
3. finally, a wrapper exhaustive search is conducted on the 2^n subsets of the n new features based on the performances of a SVM on the training set.

The advantages of using a linear model in step 2 are not only computational: linear models are immune to variable scaling issues and can handle reasonably well high-dimensional data.

The supervised dimensionality reduction method proposed in this paper for step 1 proceeds by clustering spectral variables that have a similar informative content about the class label. The clustering algorithm is therefore applied on features (*i.e.* the spectral variables) and not on spectra. The proposed method is an agglomerative bottom-up algorithm. This kind of algorithms usually requires the following elements: a measure of similarity between the features to be clustered, a measure of similarity between clusters, a fusion algorithm and a way to represent each cluster.

2.1 Similarity between spectral variables

Different criteria such as the correlation [1] and the mutual information [2] have been used to estimate the degree of similarity between spectral variables. However, the actual value of the parameter of interest (*i.e.* the class label or the value to predict) is not taken into account in these cases, which is certainly not optimal. Indeed, two features may not be as such related to each other, but their informative content regarding the parameter of interest may be the same. To overcome this problem, a supervised similarity measure is developed. It is inspired by redundancy estimation strategy developed in [3] for regression problems. The criterion proposed in [3] cannot however be applied directly for classification. Indeed, this similarity measure requires the use of distances in the joint spaces $X(j)$,Y where $X(j)$ is a spectral variable and Y the parameter of interest. In the case of classification, the parameter of interest takes discrete values only and the notion of distance in the joint space does not make sense anymore. The similarity criterion must then be adapted.

The principle of the similarity measure is the following: the two compared spectral variables $X(1)$ and $X(2)$ are similar if the number of spectra which are similar according to these features but not according to the class label is also similar. This concept of "local outliers" in the joint space $X(j)$,Y is illustrated in Fig. 1 for a toy example. The two classes are represented by crosses and triangles and all learning spectra are sorted according to feature $X(1)$. The spectra considered as similar to

spectrum *i* are its *k* nearest neighbors in the space defined by *X(1)*. Here, *k* equals 5. The local outliers are the two spectra in the neighborhood labeled with crosses.

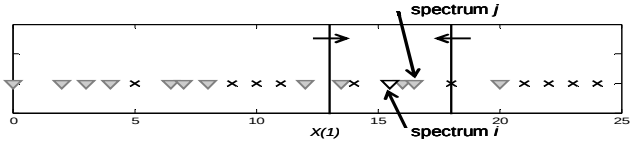

Fig. 1. The 5 neighbors of spectrum *i* with respect to the spectral variable *X(1)* are the triangles and the crosses between the two vertical lines. The crosses and the triangles represent the labels associated to the spectra. The two spectra labeled with crosses in this neighborhood are the local outliers of spectrum *i*.

A sequence with the numbers of these local outliers is built for each of the two compared spectral variables *X(1)* and *X(2)*. These sequences are respectively called n_{X1} and n_{X2}. Each element of the sequences corresponds to a spectrum in the learning set. In the illustrated example, the i^{th} element of $n_{X(1)}$ is thus 2 because two of the five nearest neighbors of spectrum *i* have the other label. The sequence $n_{X(2)}$ is built in an analogous way. The correlation between the two sequences $n_{X(1)}$ and $n_{X(2)}$ can consequently be considered as a measure of the similarity between the two features in a supervised manner. The proposed criterion of similarity between spectral variables implies a single parameter *k*, which is the number of neighbors to take into account.

2.2 Similarity between clusters

In the proposed approach the full linkage criterion is used. This means that the similarity between two clusters corresponds to the minimum similarity between each pair of elements (*i.e.* spectral variables) of the clusters. This choice ensures a maximal homogeneity of the clusters.

2.3 Fusion Algorithm

The fusion algorithm adopted for this problem has been developed in [1]. It consists of a hierarchical bottom-up algorithm for which only consecutive clusters (in the sense of the spectral variables) are compared. The restriction to consider only consecutive clusters ensures that the clustering will define ranges of consecutive wavelengths (or wavenumbers) and are therefore easier to interpret.

2.4 Representation of a Cluster

Each cluster must be represented in order to build a model based on the clustering. In this approach, the clusters are summarized by the mean of the included features. This choice corresponds to a piecewise constant approximation of the spectra.

2.5 Number of Clusters

The final number of clusters is chosen by minimizing the 3-fold cross-validation error of a discrimination model built from the cluster centroids. For this purpose, a linear model is used in order to keep the computational time reasonable and to avoid issues related to variable scaling and to high-dimensionality. This part of the proposed approach is therefore also supervised, since the target value is taken into account. While this is not the case in the experiments reported in Section 3, a filtering

approach based on the mutual information (see [4] and [5]) between the new features and the target could be applied to reduce the computational requirements.

3 Experiments

This section describes the datasets used to evaluate the efficiency of the proposed method. The experimental methodology and the results are presented.

3.1 Data and experimental methodology

The proposed supervised clustering approach is illustrated on two datasets from the food industry. The first one is the Tecator [6] which consists of 215 near-infrared spectra of meat samples. The spectra are recorded between 850 and 1050 nm and are discretized into 100 spectral variables. Two classes are defined on the dataset. The first class consists of all the spectra with less than 14% of fat and the second class includes the other spectra. The two classes count 109 spectra and 106 spectra respectively. All spectra are normalized to zero mean and unit variance and are divided into a learning set of 172 spectra and an independent test set of 43 spectra used to evaluate the performances and not to choose any parameter.

The second dataset (Wine [7]) is composed of 124 spectra of wine samples which consist of absorbance measurements recorded in the mid-infrared range at 256 different wavenumbers between 4000 and 400 cm^{-1}. Spectra number 34, 35 and 84 are considered as outliers and removed from the database. As in the previous case, two classes of spectra are built: the first one (63 spectra) corresponds to samples with an alcohol concentration smaller than 12.5 and the other one (58 spectra) includes all other samples. The dataset is divided into learning and a test set of 91 and 30 spectra.

The experimental methodology for each database is summarized in Fig. 2: the proposed approach detailed in Section 2 is compared to a SVM built on raw features and a SVM built on the centroids of a non-supervised clustering (as proposed in [1]).

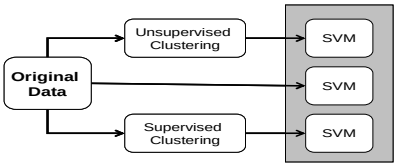

Fig. 2. Summary of the experimental methodology applied to each databases.

For both types of clustering, the number of clusters is chosen according to a 3-fold cross-validation error. The final clustering of Tecator counts 7 clusters for the supervised approach and 5 for the unsupervised methodology. In the case of the Wine dataset, the supervised and unsupervised clusterings produce 9 and 35 clusters respectively. SVM models corresponding to all possible combinations of the clusters are then built. For both clustering methods, the best subset of the new variables is chosen according to the classification error on the training set. The number k of neighbors taken into account for the supervised clustering is chosen according to the same error. This parameter equals 3 for Tecator and 6 for Wine.

3.2 Results and discussion

In the case of Tecator, the proposed supervised clustering leads to the selection of 6 clusters represented in grey on the right part of Fig. 3. The left part of the figure

shows the unique selected cluster obtained by the unsupervised clustering methodology. For the Wine database, the supervised and the unsupervised clustering approaches give respectively 5 clusters and 1 cluster of variables (Fig. 4).

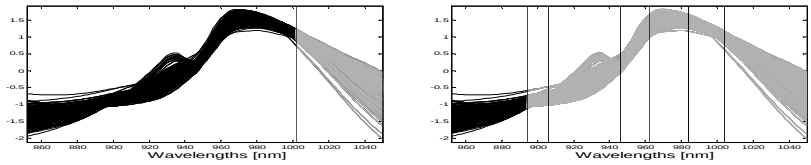

Fig. 3. Final selections of clusters for the unsupervised (left) and supervised (right) approaches on the Tecator dataset.

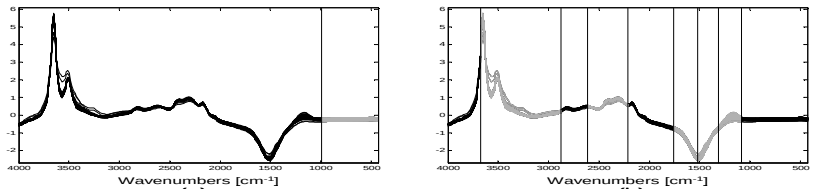

Fig. 4. Final selections of clusters for the unsupervised (left) and supervised (right) approaches on the Wine dataset.

The classification performances for the two datasets and the number of variables (either original variables or new cluster based ones) implied in the SVM models are presented in Table 1 and Table 2 for the three compared approaches. For the Tecator dataset, the best performances are achieved by the supervised clustering methodology (2.32% of incorrect classifications), which is better than the SVM built on raw features and the unsupervised clustering. Both clustering approaches help reducing drastically the number of variables included in the model. For Wine, the best performances correspond also to the supervised clustering (36.67% of incorrect classifications).

Method	Nb of (latent) variables	% of incorrect classifications
SVM on Raw features	100	4.65
NS Clustering + SVM	1	37.21
Supervised Clustering + SVM	6	2.32

Table 1. Classification performances of the three approaches for Tecator.

Method	Nb of (latent) variables	% of incorrect classifications
SVM on Raw features	256	40.00
NS Clustering + SVM	1	56.67
Supervised Clustering + SVM	5	36.67

Table 2. Classification performances of the three envisaged approaches for Wine.

In the case of Tecator, the supervised clustering approach leads to the selection a cluster including the wavelengths range around 930, which corresponds to a bump in the spectra. This result is in agreement with what can be found in [8] and [1]. It should be noted that the better performances of the supervised clustering method compared with the SVM built on the original data can be explained by the reduced number of variables implied in the classification model. Indeed, the model is not

"polluted" by variables which are not pertinent for the classifications or which are redundant with other variables.

The supervised clustering allows to select the first peak of the Wine spectra, which is in agreement with [1] and very satisfactory when analyzing the alcohol concentration. Indeed, the wavenumbers range around 3600 cm^{-1} corresponds to the absorption range of the O–H bond present in alcohol.

4 Conclusion

Reducing the dimensionality of NIR spectra is an important issue in order to avoid the difficulties related to the curse of dimensionality and to build models easier to interpret. Clustering the spectral variables allows tackling the problem by defining ranges of wavelengths. Moreover, a supervised clustering is certainly more appropriate since the target value is taken into account. Such clustering has been proposed in [3] for regression problems, but this methodology cannot be applied as such on classification problems and has to be adapted to a discrete target value.

In this paper, we propose a supervised clustering methodology with a modified similarity measure between features, which can be applied to classification problems. This approach is applied to two real world datasets and compared to models build from an unsupervised clustering and from the original features. The two type of clustering methodologies help to reduce drastically the number of variables implied in the models for the two datasets. Moreover, the resulting clusters correspond mainly to parts of the spectra identified as meaningful in the literature ([8] and [1]). The model performances are improved in both cases with the supervised clustering.

Some variations of the proposed methodology could be considered. For instance, clusters of spectral variables can be summarized by their mean values but also by additional values (e.g. maximum).

References

[1] C. Krier, F. Rossi, D. François, M. Verleysen: A data-driven functional projection approach for the selection of feature ranges in spectra with ICA or cluster analysis; Chemometrics and Intelligent Laboratory Systems, vol. 91, pp. 43-53, Elsevier (2008).

[2] G. Van Dijk, M. Van Hulle: Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis, *proceedings of the* International Conference of Artificial Neural Network, pp. 3140, Springer, Heidelberg (2006).

[3] D. François, C.Krier, F. Rossi., M. Verleysen: Estimation de redondance pour le clustering de variables spectrales, *proceedings of the* AGROSTAT 2008 10th European Symposium on Statistical Methods for the Food Industry, pp. 5561, Louvain-la-Neuve/Belgium (2008).

[4] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh, Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing), Springer (2006).

[5] V. Gomez-Verdejo, M. Verleysen, J. Fleury, Information-theoretic feature selection for the classification of hysteresis curves, *proceedings of the* IWANN'07, International Work-Conference on Artificial Neural Networks, San Sebastian (Spain). Computational and Ambiant Intelligence, Francisco Sandoval et al. eds., Springer (Berlin-Heidelberg), Lecture Notes in Computer Science 4507, pp. 522-529 (2007).

[6] Tecator meat sample dataset, http://lib.stat.cmu.edu/datasets/tecator.

[7] Dataset provided by Prof. Marc Meurens, Université catholique de Louvain, BNUT unit, meurens@bnut.ucl.ac.be. Available from http://www.ucl.ac.be/mlg/index.php?page=DataBases.

[8] F. Rossi, A. Lendassse, D. François, V. Wertz, M. Verleysen: Mutual information for the selection of relevant variables in spectrometric nonlinear modeling, Chemometrics and Intelligent Laboratory Systems, vol. 80, pp. 215226, Elsevier (2006).