ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Feature clustering and mutual information for the selection of variables in spectral data

C. Krier[1], D. François[2], F.Rossi[3], M. Verleysen[1]
[1,2] Université catholique de Louvain, Machine Learning Group
[1] DICE, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium
{krier, verleysen}@dice.ucl.ac.be
[2] CESAME Research Center, Av. G. Lemaître 4, B-1348 Louvain-la-Neuve, Belgium
françois@inma.ucl.ac.be
[3] Projet AxIS, INRIA-Rocquencourt, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France, Fabrice.Rossi@inria.fr

Abstract. Spectral data often have a large number of highly-correlated features, making feature selection both necessary and uneasy. A methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information is proposed. The clustering allows reducing the number of features to be selected by grouping similar and consecutive spectral variables together, allowing an easy interpretation. The approach is applied to two datasets related to spectroscopy data from the food industry.

1 Introduction

Many problems in spectrometry require predicting a quantitative value from measured spectra. The major issue with spectrometric data is their functional nature; they are functions discretized with a high resolution. This leads to a large number of highly-correlated features; many of which are irrelevant for the prediction. The features are furthermore ordered: similar indexes correspond to close portions of the spectra.

Features must thus be selected to reduce the complexity of the model and avoid convergence problems, overfitting, and the 'curse of dimensionality' in general. The interpretability of the selected features is of great importance for practitioners who need to understand the underlying phenomena.

The large initial number of features and the high degree of collinearity render the feature selection procedures slow and often unstable [1]. Highly–correlated features furthermore make interpretation uneasy.

This paper addresses these problems by first grouping 'similar' features together and then selecting groups of features instead of individual ones. The group of features is replaced by a 'mean' feature for selection/prediction purpose. The proposed approach furthermore enforces the interpretability of the groups by allowing only consecutive features in clusters. Hence, each cluster corresponds to a specific range of the spectra. Section 2 briefly presents the state of the art, Section 3 describes the proposed methodology and Section 4 suggests some experiments.

2 State of the art

2.1 Feature grouping

Grouping features together can be done using feature clustering or by exploiting the functional nature of the data.

Feature clustering.
Classical clustering methods, and especially hierarchical methods, are easy to transfer to cluster features instead of data elements. The only particularization that is needed is a meaningful definition of the similarity between features. See, for instance, [2] for an example of a bottom-up clustering of features on spectral data. The problem however for functional data is that the method does not necessarily select consecutive features; a cluster can thus correspond to several distinct portions of the spectra, making cluster interpretation difficult.

Functional modeling
Another approach for grouping features is to describe the spectra in a functional basis whose basis functions are 'local' in the sense that they correspond to well-defined portions of the spectra. Splines have been used successfully for this [1]. However, while the clusters always contain consecutive features, they are all forced to have the same size in [1]. Also, the contribution of each original feature to the cluster depends on its position on the wavelength range; while interpretation is possible, it is based on an approximate view of the functional features.

2.2 Feature selection

Relevance estimation
The mutual information (MI) between two random variables X and Y is defined as the uncertainty reduction on Y when X is known. In terms of probability densities, the MI can be written as

$$MI(X,Y) = \int \mu_{X,Y}(x,y) \log \frac{\mu_{X,Y}(x,y)}{\mu_X(x)\mu_Y(y)} dxdy ,$$

where μ_X and μ_Y are the probability density functions (pdf) of X and Y respectively, and $\mu_{X,Y}$ is the joint distribution of the two variables.

Feature Subset Selection
To find the optimal feature subset, several strategies can be considered. The easiest way is to score each feature and to choose the individually most relevant ones. The contrasting approach is to test all possible subsets from the set of initial features, demanding much more computations, but avoiding missing important features. A method generally accepted as a good compromise is the forward search [3]. It consists in choosing the most relevant feature first, and then finding the most relevant pair of features including the chosen first feature. The procedure is then iterated. Once a feature is chosen, it is never questioned again.

3 Proposed approach

3.1 Feature clustering

The algorithm presented here is similar to the hierarchical feature clustering algorithm [2] except that it forces every cluster to contain consecutive features only, like in the spline approach [1]. The algorithm can be characterized as an agglomerative (bottom-up) full linkage algorithm. It first merges the two most similar consecutive features

into one cluster. Feature X_j is thus compared only with X_{j-1} and X_{j+1}. The similarity between two features is estimated using the absolute value of the correlation:

$$S(X_i, X_j) = \left| \frac{E(X_i X_j) - E(X_i)E(X_j)}{\sqrt{Var(X_i)Var(X_j)}} \right|.$$

Note that for this similarity measure, the correlation is preferred (as in [4]) to the MI (as in [2]). Indeed, consecutive features are highly correlated, making the correlation as useful as the MI. Moreover, the correlation is much easier to estimate (by replacing the expectations and variances in the above equation by sums).

The algorithm then recursively merges at each step the two most similar consecutive clusters. The similarity between two clusters is defined as the minimum similarity between each element of one cluster and the elements of the other. The algorithm stops when all features have been grouped into a single cluster and provides with a dendrogram representing the clustering.

The output value associated with each cluster, its representative, is chosen to be the mean of the spectra over the range of features defined by the cluster. The spectra are thus approximated by a piecewise constant function; the original features are replaced with the (fewer) representatives of each clusters. The number of clusters is chosen to maximize the prediction performances of a linear model built on the feature clusters. The efficiency of the linear models is measured by the Normalized Mean Squared Error (NMSE) in the same cross-validation method used to build the predictive model on the selected features (for practical details, see Section 4). This NMSE is defined as follows:

$$NMSE_{\Omega} = \frac{\frac{1}{N_{\Omega}} \sum_{\Omega} \left(\hat{Y}_i - Y_i \right)^2}{Var(Y)},$$

where Y_i is the parameter to predict, $\hat{Y}_i$ the corresponding prediction and N_{Ω} the number of data in Ω.

3.2 **Feature cluster selection**

Once the clusters have been constructed, the most relevant ones for the prediction are selected. The selection is performed using the MI with a procedure similar to [3]. To this aim, a forward feature selection is performed, as described in the previous section about Feature Subset Selection. The most relevant group of features identified by the method is then kept. The next step consists in an exhaustive search over this subset: if the group includes K features, all 2^K-1 not-empty combinations of these features are tested, in a wrapper procedure using the prediction model itself. The model used here is a Radial Basis Function Network (RBFN).

In theory, the forward search with the mutual information criterion should never stop: the mutual information cannot decrease when variables are added in the set. However, in practice when using a MI estimator, this is not true anymore (the quality of the estimator decreases when the dimensionality becomes high). The selection procedure is thus stopped when the MI estimation decreases. In addition, to keep the

computation time of the subsequent exhaustive search within reasonable limits, the forward search is stopped in any case when a threshold number of features is reached.

4 Experiments

The method proposed in this paper is evaluated on two datasets. The first one is the Tecator database [5]. It consists of 215 near infrared spectra of meat samples, in the 850–1050 nm wavelength range. Each spectrum is discretized into 100 spectral variables. The parameter to predict is the fat content of the meat samples. The learning set contains 172 spectra, which leaves 43 spectra for the test.

The second dataset, Wine [6], consists of mid-infrared spectra of wine samples recorded at 256 wavenumbers, from which alcohol concentration has to be predicted. The learning and test sets contain 94 and 30 spectra respectively; three spectra (# 34, 35 and 84) are considered as outliers and therefore discarded.

The approaches

Three approaches are compared for each dataset: the first one is the clustering approach described above. The second one consists in building a RBFN on all spectral variables. The last one is a projection on a B-Spline basis, followed by the same selection procedure as the one in the clustering approach. The optimal meta-parameters number of units and variances) of each RBFN are chosen by grid search in a 3-fold cross-validation procedure over the learning set.

Results

Figure 1 shows in grey the portion of the spectra selected by the clustering approach for the Tecator dataset. The vertical lines indicate the limits of the four selected clusters. The NMSE on the test set equals 0.1342 for the model build on all spectral variables; in the case of the spline approach, the NMSE is 0.1059 (with 30 splines). Concerning the clustering approach, the NMSE is 0.1209. For comparison, the NMSE of a cmodel predicting the mean of the parameter of interest would be 1.

The portions of spectra corresponding to selected clusters are presented in grey in Figure 2 for the Wine database. The selected clusters are delimited by the vertical lines. The NMSE in test reaches 0.0802 for the proposed methodology, 0.0938 for the spline approach and 0.1254 for the model built on all variables.

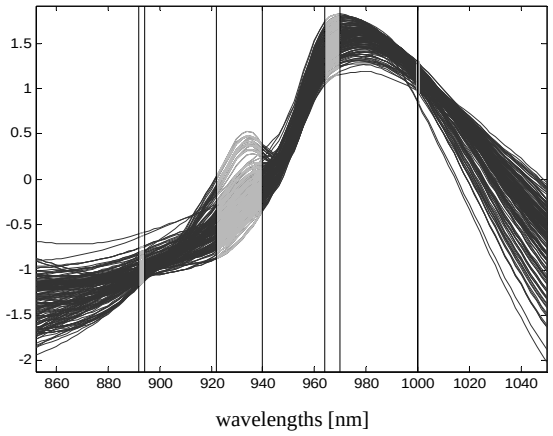

Fig. 1: Tecator dataset: Parts of spectra corresponding to selected clusters (in grey).

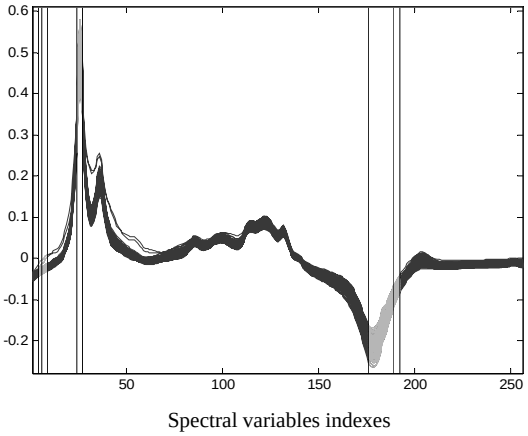

Fig. 2: Wine dataset: Parts of spectra corresponding to selected clusters (in grey).

Discussion

From Figure 1, it appears that the second cluster (922 to 940 nm) contains 3 of the 7 wavelengths selected by a non-functional MI selection as described in [7]. This highlights the redundancy between these spectral variables. The first cluster corresponds approximately to the 4 other wavelengths selected in [7]. The two other clusters are different. In the case of Wine, the five selected clusters (see Figure 2) cover approximately 8 out of the 20 spectral variables selected in [8], [9].

For the two datasets, the model built on all features has the worst performances, while the clustering approach leads to better predictions. This last one performs also better than the spline methodology, except for the Tecator dataset. The spectra in this dataset are indeed much smoother than those in Wine, and an approach by spline projection has been shown to be efficient (see [1]). However, the Wine dataset seems to benefit from a clustering approach more adapted to the particularities of the data.

5 Conclusions

This paper proposes a methodology to build efficient prediction models on functional data, for example smooth spectra. The model groups consecutive features in clusters, and replaces each of the latter by a single feature. Such kind of feature selection allows the interpretation of the selected variables, contrarily to variable projection methods where this interpretation is lost. Compared to previous approaches of feature clustering, the proposed one uses the functional nature of the data to force only consecutive features to be grouped together in a cluster. Compared to functional methods, the proposed approach has the advantage that the clusters are not forced to be of equal size, does not weight the features according to their position in the cluster, and effectively builds the clusters according to the specificities of the data.

The models obtained by this methodology perform better than models built on all spectral variables in the two spectroscopic benchmarks presented in this paper. Moreover, the low number of clusters identified by the method allows the interpretation of the selected variables: several of the selected clusters include the spectral variables identified on these benchmarks as meaningful in the literature.

6 Acknowledgments

C. Krier is funded by a Belgian FRIA grant. Parts of this research results from the Belgian Program on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with its authors.

7 References

[1] F. Rossi, D. François, V. Wertz, M. Verleysen, *Fast Selection of Spectral Variables with B-Spline Compression*, Chemometrics and Intelligent Laboratory Systems, Elsevier, in press.

[2] G. Van Dijk, M.M. Van Hulle, *Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis*, International Conference on Artificial Neural Networks 2006, Athens, Greece.

[3] C. Krier, D. François, V.Wertz, M. Verleysen, Feature Scoring by Mutual Information for Classification of Mass Spectra, FLINS 2006, 7th International FLINS Conference on Applied Artificial Intelligence, August 29-31, 2006, Genova (Italy).

[4] T. Lan, D. Erdogmus, M. Pavel, S. Mathan, *Automatic Frequency Bands Segmentation Using Statistical Similarity for Power Spectrum Density Based Brain Computer Interfaces,*Proceedings of IJCNN 2006, Vancouver, 2006.

[5] *Tecator meat sample dataset*. Available on Statlib: http://lib.stat.cmu. edu/datasets/tecator.

[6] Dataset provided by Prof. Marc Meurens, Université catholique de Louvain, BNUT unit, meurens@bnut.ucl.ac.be. Dataset available from http://www.ucl.ac.be/mlg/.

[7] F. Rossi, A. Lendasse, D. Francois, V. Wertz and M. Verleysen, *Mutual information for the selection of relevant variables in spectrometric nonlinear modeling*, Chemometrics and Intelligent Laboratory Systems 80, pages 215 – 226, Elsevier, 2006.

[8] N. Benoudjit, E. Cools, M. Meurens and M. Verleysen, *Chemometric calibration of infrared spectrometers: selection and validation of variables by non-linear models*, Chemometrics and Intelligent Laboratory Systems 70, pages 47– 53, Elsevier, 2004.

[9] N. Benoudjit, Variable selection and neural networks for high-dimensional data analysis, Application in infrared spectroscopy and chemometrics, PhD Thesis, Université Catholique de Louvain, November 2003.