Estimation de redondance pour le clustering de variables spectrales

Estimating redundancy for spectral variable clustering

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

[1] *DMF-Solutions & UCL, Machine Learning Group, av. G. Lemaître, 4, B-1348 Louvain-la-Neuve*
E-mail : damien.francois@dmf-solutions.be

[2] *UCL, Machine Learning Group, Place. du Levant, , B-1348 Louvain-la-Neuve*
E-mail : catherine.krier@uclouvain.be, *michel.verleysen@uclouvain.be*

[3] *Projet AxIS, INRIA-Rocquencourt, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France,*
E-mail : Fabrice.Rossi@inria.fr

Résumé
La réduction du nombre de variables spectrales dans un problème de modélisation permet souvent de construire un modèle plus simple, plus performant, et apporte en sus une information sur les 'fréquences utiles' à la prédiction, auxquelles on peut trouver une interprétation. Ce papier propose une approche pour la sélection de bandes de fréquences (variables spectrales consécutives) à l'aide d'un clustering hiérarchique qui prend en compte la variable à prédire pour effectuer le regroupement des variables. La méthode est basée sur l'estimateur par plus proche voisins de l'information mutuelle. La méthode est illustrée sur deux jeux de données réels.
Mots-clés : Spectroscopie, sélection de variables, clustering de variables, information mutuelle, PLSR.

Abstract
Reducing the number of spectral variables in a calibration problem often allows building simpler, and more accurate, models. It furthermore brings information about the relevant frequency ranges, for which an interpretation can be sought. This paper proposes a hierarchical clustering approach, which takes into account the target variable, to merge feature clusters. The method is based on the nearest neighbour estimator of mutual information. The method is illustrated on two real datasets.
Keywords : Spectroscopy, variable selection, variable clustering, mutual information, PLSR

1. Introduction

De manière générale, les variables spectrales, décrivant des spectres lisses (comme par exemple le spectres infra-rouges), étant une discrétisation par exemple de l'espace continu des longueurs d'ondes, sont fortement colinéaires, et par conséquent fort redondantes. Cette redondance entraîne des modèles inutilement complexes, sujets au sur-apprentissage, à des problèmes de convergence, etc.

Une approche souvent envisagée pour éviter ces écueils est de réduire le nombre de variables en projetant linéairement les variables sur un sous-espace de dimension inférieure dont les axes sont décorrélés (PCR, PLSR, etc.).

Cette approche, bien que réduisant la complexité du modèle, résulte toujours de l'utilisation de l'ensemble des variables de départ, et l'interprétation du modèle en terme de longueur d'ondes pertinentes est délicat et demande beaucoup de travail et d'expérience.

Le clustering des variables spectrales permet de réduire la complexité des modèles de prédictions, mais fournit en plus une information très facile à interpréter en termes d'intervalles de longueur d'ondes. Les variables sont regroupées en ensembles homogènes quant à leur 'contenu d'information'.

Cette communication présente un algorithme de clustering hiérarchique adapté aux variables spectrales, et deux critères d'estimation de redondance entre variables spectrales. Elle est organisée comme suit: la section 2 présente la notion de clustering de variables, la section 3 propose deux critères d'estimation de redondance entre variables, la section 4 illustre ces concepts sur deux bases de données, et la section 5 conclut.

2. Clustering 'bottom-up' des variables spectrales

Les algorithmes de clustering hiérarchiques fonctionnent itérativement en regroupant à chaque étape les clusters les plus similaires (bottom-up, ou agglomératif), ou en scindant les groupes les moins homogènes (top-down, divisive). Les algorithmes agglomératifs sont tous basés sur la combinaison des éléments suivants (Kaufmann, 1990) :

- un critère pour estimer la similarité entre éléments ;
- une méthode pour estimer la similarité entre clusters ;
- un algorithme de fusion de clusters ;
- une méthode pour choisir un représentant d'un cluster. ;

Par exemple, pour le clustering d'observations, on utilise typiquement :

- la distance euclidienne entre clusters singletons ;
- le full linkage, single linkage, centroid linkage, ou average likage entre clusters ;
- un algorithme de fusion qui consiste à prendre les deux clusters qui sont les plus semblables et à les fusionner dans un même cluster, puis à itérer ;
- le centroide, ou le médoïde comme représentant de cluster.

2.1. Adaptation au clustering de variables

Cette méthodologie peut être appliquée également aux variables ; plutôt que de regrouper les spectres les plus similaires, on va regrouper les variables spectrales les plus similaires. Cela demande cependant quelques adaptations, détaillées ci dessous.

2.1.1. Estimation de la similarité entre éléments

Les éléments sont ici les variables. Le première critère qui vient à l'esprit pour estimer la 'similarité' entre variables est la (valeur absolue de la) corrélation. Plus la corrélation est élevée, plus les variables 'apportent la même information'. La section suivante proposera un autre critère.

2.1.2. Estimation de la similarité entre clusters

De même que pour le clustering d'observations, les critères de similarité entre clusters de variables peuvent être définis à partir du critère de similarité entre variables seules. Le 'single linkage' consisterait à choisir le maximum de corrélation entre chaque paire de variable d'un cluster et de l'autre. Le full linkage correspond au minimum de corrélation. Le 'average linkage' correspondrait à faire la moyenne des corrélations, et le 'centroid linkage' à calculer la corrélation des moyennes des variables spectrales.

Il est à noter cependant que le critère de similarité entre clusters peut se construire indépendamment des critères de similarité individuelles, comme on le verra dans la section suivante.

2.1.3. Algorithme de fusion

Bien que l'algorithme classique puisse être utilisé sans problèmes et donner des résultats intéressants, (Van Dijk, 2006), il ne tient pas compte de l'ordre naturel des variables et pourra regrouper ensemble des variables fort éloignées dans le spectre. Il est cependant fort simple de l'adapter de manière à ce qu'il ne fusionne que des variables ou clusters de variables consécutifs. Ainsi, le clustering résultant ne fera apparaître que des groupes contigus de variables, chaque cluster étant identifié à un sous intervalle de longueurs d'ondes (Krier, 2007).

2.1.4. Choix du représentant d'un cluster

Le choix du représentant du cluster, important pour la construction du modèle de prédiction subséquent, peut se faire de diverses manières : valeur moyenne des variables, valeur maximum des variables, valeur correspondant à la variable la plus similaire à la valeur à prédire, etc. Le choix de la moyenne revient à approximer les spectres par des fonctions constantes par morceaux. D'autres choix correspondent à d'autres approximations des spectres (par exemples des splines),

2.2. Choix du nombre de clusters

La procédure de clustering fournit un arbre (dendrogramme) de tous les clusterings possibles de 1 à p clusters, où p est le nombre de variables spectrales. Le choix effectif du nombre de clusters peut se faire, dans notre application, en minimisant l'erreur de cross-validation du modèle de prédiction construit en utilisant comme variables les représentants de chaque cluster.

3. Estimation de redondance

Bien que la corrélation soit un bon moyen de comparer les contenus d'information de deux variables, surtout lorsqu'elle sont à priori linéairement dépendantes, elle a pour inconvénient qu'elle ne tient pas compte de la variable à prédire, et mène donc à un clustering des variables qui est purement indépendant de la variable cible.

Pour pallier à ce défaut, et permettre de comparer le contenu informationnel de deux variables par rapport à une variable cible Y, on peut se tourner vers une approche par plus proches voisins (François 2007).

Le principe est le suivant. Pour mesurer la similarité du contenu d'information entre deux variables spectrales X1 et X2, deux séries de valeurs $n_{X1.}$et $n_{X2.}$ sont construites. Chaque i^{eme} élément de n_{X1} est calculé comme suit, par rapport au i^{eme} spectre de l'ensemble de calibration.

1. On trouve le spectre le plus proche du i^{eme} au sens de la distance euclidienne sur l'espace joint (X1, Y), c'est à dire pour lequel la valeur de X1 et de Y correspondant sont les plus proches. Supposons que ce soit le j^{eme}.

2. On compte le nombre de spectres dans l'ensemble de calibration pour lesquels on ait à la fois que la valeur en X1 est plus proche du i^{eme} spectre que le j^{eme}, mais que la variable en Y soit plus différente que celle associée au j^{eme} spectre. On compte ainsi les spectres qui sont similaires au i^{eme} mais qui correspondent à des valeurs cibles fort différentes, ce sont des 'voisins mal placés'. Si ce nombre est grand, le i^{eme} spectre peut être interprétés comme un 'outlier local' qui s'écarte de la relation fonctionnelle entre les variables d'entrées et la variable à prédire.

La seconde série n_{X2} est construite de manière semblable pour X2.

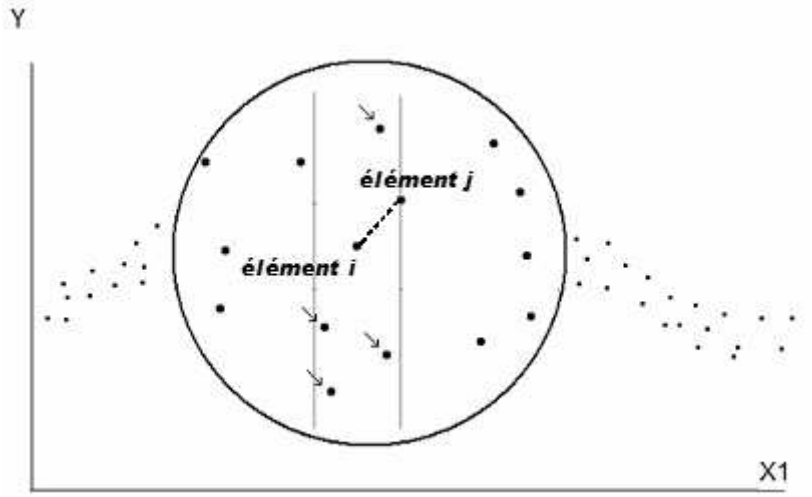

Figure 1. Graphe de Y = f(X1); agrandissement sur la partie centrale. L'élément i est le spectre courant, et l'élément j son plus proche voisin en termes de X1 et de Y. Les points indiqués par une flèche sont les 'voisins mal placés' (voir texte). Leur nombre est la $i^{ème}$ valeur de n_{X1}.

Si deux variables, pouvant être différentes, apportent la même information par rapport à Y, alors ce sont les mêmes spectres qui seront 'outlier locaux'. Deux modèles de prédiction construits chacun sur une de ces variables feront des erreur corrélées, et se 'tromperont' pour les mêmes.

La corrélation entre n_{X1} et n_{X2} est donc une mesure de la similarité d'information que X1 et X2 apportent par rapport à Y, la variable à prédire. La valeur corr(n_{X1}, n_{X2}) servira de mesure de similarité entre variables dans l'algorithme de clustering. Il est important de noter que, contrairement à la corrélation, la mesure de similarité définie ci-dessus s'applique aussi bien à des variables qu'à des groupes de variables.

Dans la section suivante, cette mesure de similarité est utilisée pour le clustering, en comparaison avec la simple corrélation, sur deux jeux de données réelles.

4. Expériences

Le jeu de données Wine [1]consiste en 124 spectres (dont 3 outliers éliminés et 30 spectres réservés pour l'ensemble de test) proche infra-rouge d'échantillons de vin pour lesquels la concentration en alcool doit être prédite. Pour le jeu de données Tecator[2], le problème consiste à prédire le taux de graisse dans des échantillons de viande, à partir de leurs spectres mesures en NIR, entre 850 et 1050 nm. Au total, 215 spectres sont disponibles ; les 150 premiers sont choisis comme base d'apprentissage, le reste étant réservé pour le test.

Pour chaque jeu de données, un clustering a été effectué d'une part sans utiliser l'information de la variable à prédire et d'autre part en suivant l'approche proposée dans ce papier. Un modèle PLS est alors construit sur les clusters obtenus. Le nombre de clusters est choisi selon un critère AIC et le nombre de composantes dans le modèle PLS par 4-fold cross-validation. A titre de comparaison, les résultats obtenus par un modèle PLS sur l'ensemble des variables spectrales (sans clustering), sont également fournis.

Les résultats (erreurs normalisées NMSE sur l'ensemble de test) sont repris dans le Tableau 1, ainsi que le nombre de clusters et le nombre de composantes PLS optimaux. Le clustering permet d'améliorer légèrement les performances par rapport au modèle construit sur toutes les variables. Dans le cas de Wine, le clustering tenant compte de la variable Y est nettement meilleur que celui obtenu sans tenir compte de la variable à prédire. Dans le cas de Tecator, l'avantage en termes d'erreur est moins marqué, mais le modèle PLS optimal est plus simple dans le cas où Y est utilisé pour le clustering (8 clusters plutôt que 17 et 7 composantes PLS plutôt que 10). Dans tous les cas, le clustering permet de réduire la complexité du modèle et permet d'obtenir de meilleurs résultats en prédiction que les modèles construits sur l'ensemble des variables d'origine; de plus, la prise en compte de la variable à prédire dans l'étape de clustering permet de réduire la complexité (donc augmenter l'interprétabilité) du modèle, tout en accroissant les performances de prédiction.

	PLSR (spectres complets)	PLSR (clustering sans Y)	PLSR (clustering avec Y)
Wine	**0.00578** 256 variables 10 facteurs	**0.0111** 19 clusters 11 facteurs	**0.00546** 33 clusters 10 facteurs
Tecator	**0.02658** 100 variables 10 facteurs	**0.02574** 17 clusters 10 facteurs	**0.02550** 8 clusters 7 facteurs

Tableau 1. Récapitulatif des résultats. Le clustering permet de réduire la complexité du modèle et permet d'obtenir de meilleurs résultats en prédiction que les modèles construits sur l'ensemble des variables d'origine.

[1] Wine alcool NIR Dataset fourni par le Prof. Marc Meurens, Université catholique de Louvain, BNUT unit, meurens@bnut.ucl.ac.be. Les données sont accessibles via http://www.ucl.ac.be/mlg/.

[2] Tecator meat sample dataset. Accessible sur statlib : http:// lib.stat.cmu.edu/datasets/tecator.

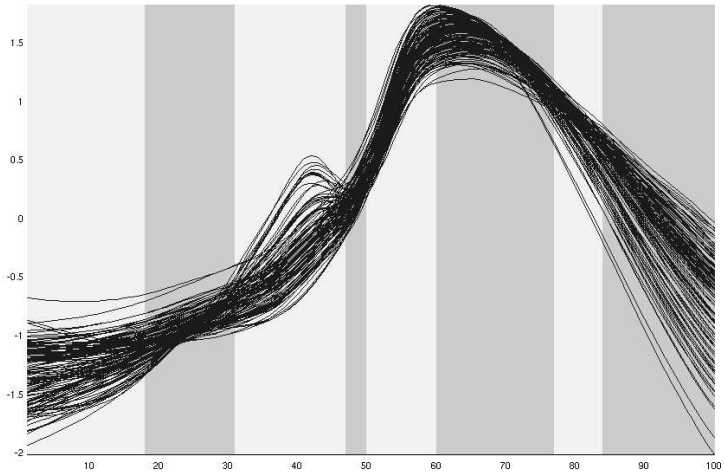

Figure 2. Clustering obtenu sur Tecator. Le premier cluster est représenté par un arrière plan clair, le second, un arrière plan foncé, le troisième de nouveau sur un arrière plan clair, etc.

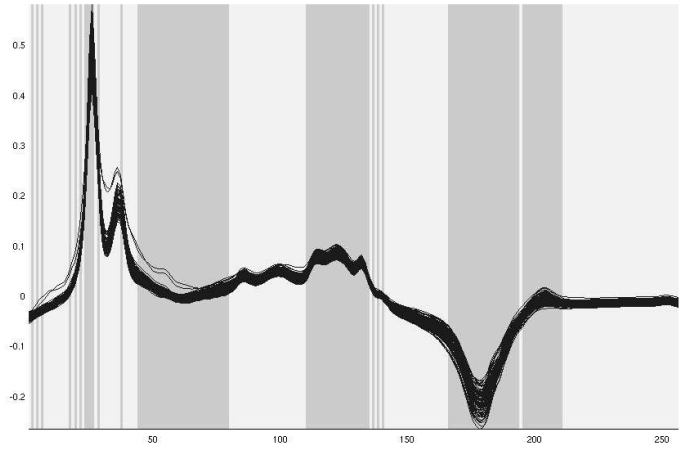

Figure 3. Clustering obtenu sur Wine. Le premier cluster est représenté par un arrière plan clair, le second, un arrière plan foncé, etc**.**

5. Conclusions

Le clustering de variables permet de regrouper en groupes homogènes les variables spectrales ayant le même contenu d'information. Outre l'information que cela peut apporter quant à l'identification des fréquences ou longueurs d'ondes d'intérêt pour l'application, le clustering permet de réduire le nombre de variables et par conséquent de construire des modèles plus simples, ayant pourtant des performances de prédiction similaires voire meilleures.

Les algorithmes de clustering traditionnels peuvent être adaptés sans problème pour le clustering de variables plutôt que le clustering d'observations, pour autant qu'une mesure de similarité de contenu entre variables soit définie. Cette mesure peut être la corrélation, qui ne tient pas compte de la variable cible, ou un critère basé sur les plus proches voisins, qui lui prendra en compte l'information de la variable à prédire.

Une fois le clustering effectué, les modèles de prédictions habituels peuvent être utilisé, avec ou sans sélection des clusters par des méthodes de sélection de variables (Rossi, 2006).

Par ailleurs, le choix du représentant des clusters peut s'avérer crucial; utiliser, par exemple le maximum des variables spectrales d'un cluster, plutôt que leur moyenne, permettrait d'atténuer les problèmes de décalage de spectres pour lesquels l'information se trouve dans les 'pics'.

Bibliographie

FRANCOIS, D, KRIER, C, ROSSI, F, et VERLEYSEN, M. (2007) 'Estimation de redondance conditionnelle par plus proches voisins, application au clustering de variables spectrales. Accepté pour publication à Chimiométrie 2007, Lyon, 29 et 30 novembre 2007.

L KAUFMANN and P J ROUSSEEUW (1990) Finding Groups in Data: An Introduction to Cluster Analysis John Wiley & Sons Ltd., Chinchester, New York, Weinheim.

KRIER C., FRANÇOIS D., ROSSI F., VERLEYSEN M., (2007) "Feature clustering and mutual information for the selection of variables in spectral data." *ESANN 2007, European Symposium on Artificial Neural Networks,* Bruges (Belgium), 25-27 April 2007, pp. 157-162.

ROSSI F., LENDASSE A., FRANÇOIS D., WERTZ V., VERLEYSEN M., (2006) "Mutual information for the selection of relevant variables in spectrometric nonlinear modeling", *Chemometrics and Intelligent Laboratory Systems*, Elsevier, Vol. 80, No. 2 (February 2006), pp. 215-226.

VAN DIJK, G., VAN HULLE, M.M. (2006) "Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis". in S. KOLLIAS ET AL. (Eds.), *Conference on Artificial Neural Networks (Europe) 2006, Proceedings of the Conference,* 10-14 September 2006, Athens, Greece, Springer-Verlag, Berlin Heidelberg, pp. 31 – 40.