Estimation de redondance conditionnelle par plus proches voisins, application au clustering de variables spectrales

Damien François[a], Catherine Krier[b], Fabrice Rossi[c], Michel Verleysen[b]

[a] *Université catholique de Louvain, Machine Learning Group, av. G. Lemaître, 4, 1348 Louvain la Neuve, Belgique, francois@inma.ucl.ac.be*
[b] *Université catholique de Louvain, Machine Learning Group, pl. du Levant, 3, 1348 Louvain la Neuve, Belgique, {krier, verleysen}@dice.ucl.ac.be*
[c] *Projet AxIS, INRIA-Rocquencourt, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France, 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. Des expériences sur deux jeux de données montrent l'intérêt de la méthode, à la fois par rapport à la PLS construite sur toutes les variables et par rapport à un clustering des variables qui ne tiendrait pas compte de la variable à prédire.

MOTS-CLÉS : *Spectroscopie, sélection de variables, clustering de variables, information mutuelle, PLSR.*

1. Introduction

Il est communément admis que les variables spectrales, de manière générale, et de spectres en proche infrarouge en particulier, sont fortement colinéaires et donc redondantes. Pour réduire cette redondance, une approche souvent envisagée est de projeter les spectres sur un sous-espace de dimension inférieure dont les axes sont décorrélés (PCR, PLSR, etc.). L'interprétabilité de ce types de modèle est cependant réduite dans le sens où les nouvelles variables construites dépendent toutes de chacune des variables initiales.

D'autre approches permettant de créer de nouvelles variables correspondant précisément à des plages spécifiques de longueurs d'ondes ont été proposées [KRIER 07, VAND 06], notamment grâce à des algorithmes de clustering regroupant les variables 'similaires', c'est-à-dire au même contenu d'information. Le clustering de variables permet de construire des modèles moins complexes (basés sur moins de variables) et interprétables directement en termes de longueurs d'onde 'pertinentes'.

Malheureusement, la majorité des méthodes de clustering ne tiennent pas compte de la variable à prédire, ou variable dépendante, dans l'algorithme de clustering. L'objectif de cette contribution est d'aller un pas plus loin en intégrant l'information de la variable à prédire dans le critère même de similarité entre variables.

2. Clustering 'bottom-up' de variables spectrales

Le clustering 'bottom-up' de variables, tel que décrit dans [VAND 06], et adapté aux données spectrales [KRIER 07], se base sur plusieurs éléments : un algorithme de fusion de groupes de variables, une fonction de mesure de similarité entre variables, et entre groupes de variables, et une méthode pour construire un représentant de chaque cluster. L'approche proposée dans [KRIER 07] est basée sur un algorithme qui ne fusionne que des groupes de variables consécutives. La similarité entre variables est estimée par corrélation, et la similarité entre groupes de variables est prise comme le maximum de similarité entre chaque paire de variables des deux groupes (full linkage). Le représentant d'un cluster est la moyenne des variables de ce cluster.

L'objectif de ce papier est de modifier cette approche de manière à tenir compte de la variable à prédire. Des efforts ont été entrepris dans ce sens, par exemple en modifiant l'algorithme de fusion de groupes [HONK 07]. Nous proposons d'intégrer l'information de la variable à prédire directement dans le critère de mesure de similarité entre variables.
Beaucoup de méthodes permettent en effet d'estimer le contenu d'information d'une variable pour en prédire une autre ; corrélation, information mutuelle, etc, mais ne permettent pas de savoir si deux variables sont redondantes conditionnellement à une troisième variable, dans notre cas la variable à prédire.

Ne pas tenir compte de la variable à prédire peut poser problème. Dans le cas où l'information nécessaire à la prédiction est par exemple cachée dans la différence entre deux variables fortement corrélées (autrement dit dans la dérivée du spectre), un clustering ignorant la variable à prédire risque de regrouper ces deux variables dans le même cluster, détruisant ainsi l'information utile.

3. Estimation de redondance conditionnelle

L'objectif est donc de comparer le contenu d'information, et pas seulement la quantité d'information, que possèdent deux variables X1 et X2 pour en prédire une troisième Y. L'approche proposée se base sur le principe de l'estimateur d'information mutuelle par plus proches voisins [KRA 04], étendu ici pour mesurer la redondance de deux variables, conditionnellement à celle à prédire.

Le principe est le suivant. Les spectres sont tout d'abord, de manière traditionnelle, séparés en ensemble d'apprentissage et de test. Pour mesurer la similarité entre les contenus d'information des deux variables X1 et X2, nous construisons deux séries de valeurs n_{X1} et n_{X2}. Chaque $i^{ème}$ élément de la série n_{X1} correspond à un élément de l'ensemble de calibration et est calculé de la manière suivante. Premièrement, on recherche dans l'ensemble d'apprentissage le spectre j le plus similaire au spectre i, la similarité étant mesurée par la distance mesurée sur les deux variables X1 et Y. Deuxièmement, on compte le nombre de spectres qui sont davantage semblables au spectre i que le spectre j lorsqu'on mesure la similarité selon X1, et qui sont en même temps moins semblables au spectre i que le spectre j lorsque la similarité est mesurée en Y. La Figure 1 montre les spectres (repérés par une flèche) qui sont dans ce cas. Il s'agit donc de 'voisins mal placés' qui sont similaires en termes de variables spectrales mais pas de valeur à prédire. Le nombre de ces spectres constitue la $i^{ème}$ entrée de la série n_{X1}. La seconde série n_{X2} est construite de la même manière, en remplaçant X1 par X2.

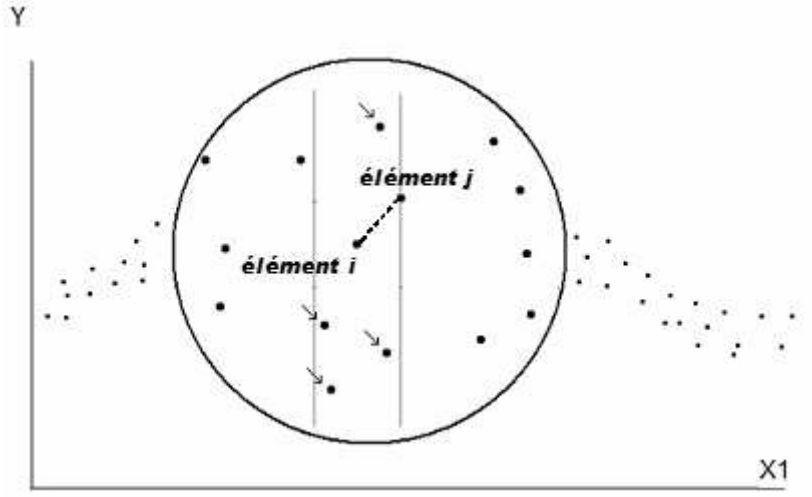

Figure 1. *Graphe de Y = f(X1); agrandissement sur la partie centrale. Le point bleu est le plus proche du point rouge ($i^{ème}$ spectre) en en tenant compte de X1 et Y. Les points indiqués par une flèche sont les 'voisins mal placés' (voir texte). Leur nombre est la $i^{ème}$ valeur de n_{X1}.*

Deux variables apportant la même information par rapport à Y auront des 'voisins mal placés' aux mêmes endroits ; ce seront les mêmes points qui sont proches en termes de variables spectrales mais éloignés en termes de valeur à prédire. La corrélation entre ces deux séries indique donc à quel point ces deux variables fournissent une information permettant de fournir correctement une prédiction pour les mêmes spectres et se tromper sur les mêmes autres, et mesure ainsi le contenu d'information commun qu'elles apportent pour prédire Y. La valeur corr(n_{X1} , n_{X2}) servira donc de mesure de similarité entre variables dans l'algorithme de clustering. Il est important de noter que la notion de corrélation décrite ci-dessus peut être directement étendue pour tenir compte de groupes de variables plutôt que de variables individuelles, ce qui est le cas dès la deuxième étape de l'algorithme de clustering hiérarchique

4. Expériences

Le jeu de données Wine [MEURENS] 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 [TECATOR], 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 [KRIER 07] 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 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 ou 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.*

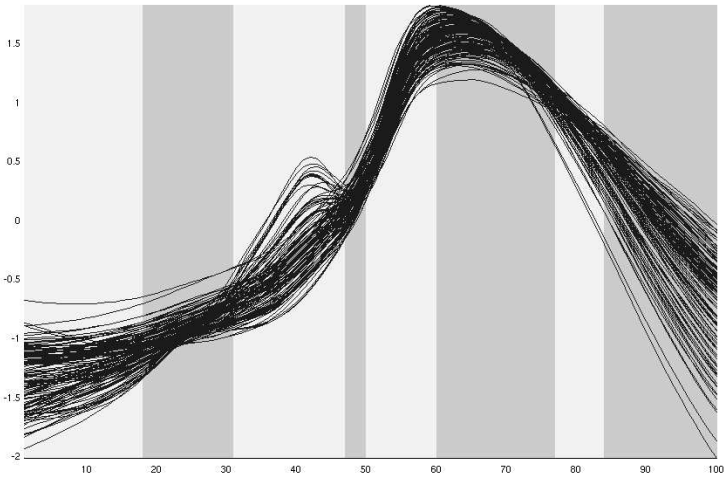

Figure 2. *Clustering obtenu sur Tecator. Le premier cluster est représenté par un arrière plan clair, le second, un arrière plan foncé, le troisième de nouveau sur un arrière plan clair, etc.*

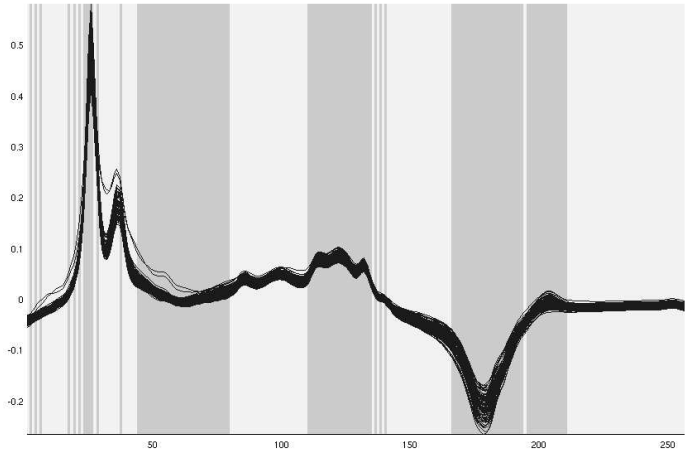

Figure 3. *Clustering obtenu sur Wine. Le premier cluster est représenté par un arrière plan clair, le second, un arrière plan foncé, etc.*

5. Conclusions

Le clustering de variables permet de regrouper ensemble des variables spectrales ayant le même contenu d'information, apportant une information visuelle sur les bandes de fréquences impliquées dans la prédiction. Bien que le clustering puisse se faire sans utiliser l'information de la variable à prédire, inclure cette information dans le critère de similarité entre variables spectrales, par une approche de plus proche voisins ayant un lien étroit avec la notion d'information mutuelle, permet d'améliorer les performances de modèles linéaires construits sur les clusters.

Il est à noter qu'une fois le clustering effectué, des méthodes de sélection de variables telles que celles présentées dans [ROS 06] pourraient être utilisées pour ne choisir que les clusters pertinents et construire un modèle de prédiction non linéaire. 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'.

Remerciements

Le travail de C. Krier est financé par une bourse du Fond pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture (FRIA, Belgique).

Bibliographie

[HONK 07] HONKELA, A., SEPPÄ, J., ALHONIEMI, E. "Agglomerative Independent Variable Group Analysis". In *Proceedings of the 15th European Symposium on Artificial Neural Networks (ESANN 2007),* Bruges, Belgium. pp 55-60.

[KRA 04] KRASKOV A., STÖGBAUER H. AND GRASSBERGER P, 'Estimating mutual information', *Phys. Rev. E69*:066138, 2004.

[KRIER 07] KRIER C., FRANÇOIS D., ROSSI F., VERLEYSEN M., "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.

[MEURENS] 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/.

[ROS 06] ROSSI F., LENDASSE A., FRANÇOIS D., WERTZ V., VERLEYSEN M., "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.

[TECATOR] Tecator meat sample dataset. Available on statlib : http:// lib.stat.cmu.edu/datasets/tecator.

[VAND 06] VAN DIJK, G., VAN HULLE, M.M. . "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, 2006, pp. 31 – 40.