Sélection de variables spectrales par information mutuelle multivariée pour la construction de modèles non-linéaires

Amaury **Lendasse**[1], Damien **François**[2], Fabrice **Rossi**[3], Vincent **Wertz**[2], Michel **Verleysen**[41]

[1] *Helsinki University of Technology – Lab. Computer and Information Science, Neural Networks Research Centre, P.O. Box 5400, FIN-02015 HUT, Finlande,* lendasse@hut.fi

[2] *Université catholique de Louvain – Machine Learning Group, CESAME, 4 av. G. Lemaître, 1348 Louvain-la-Neuve, Belgique,* francois@auto.ucl.ac.be

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

[4] *Université catholique de Louvain – Machine Learning Group, DICE, 3 place du Levant, 1348 Louvain-la-Neuve, Belgique,* verleysen@dice.ucl.ac.be

MOTS CLÉS : Spectroscopie, modèles non-linéaires multivariés, sélection de variables, information mutuelle

1. Introduction

De nombreux problèmes analytiques liés à la spectrométrie requièrent la prédiction d'une variable quantitative en fonction des variables spectrales mesurées; par exemple, on peut chercher à prédire la concentration d'un composant dans un produit dont le spectre a été mesuré.

Des modèles linéaires de prédiction, tels que la régression sur composantes principales (PCR), la régression en moindres carrés partiels (PLSR) ou la régression linéaire multiple pas-à-pas (SMLR) sont couramment utilisés pour résoudre ce type de problème [BER 00], [MAS 97]. Ils ont l'avantage d'être éprouvés et aisés à utiliser; par exemple, ils ne nécessitent généralement pas de choix de paramètres par l'utilisateur. Dans certains cas cependant, la relation physique entre données spectrales et variable à prédire ne peut pas être approchée de façon linéaire [BER 99]. L'utilisation de modèles non-linéaires devient alors indispensable.

Un spectre mesuré comprend un nombre important (plusieurs centaines) de variables. Il est nécessaire d'en sélectionner un sous-ensemble à utiliser par les modèles linéaires, pour éviter les problèmes de co-linéarité [EKL 99]; c'est ce que font des méthodes telles que PCR, PLSR ou SMLR. Dans le cas de modèles non-linéaires, le problème devient encore plus crucial: trop de variables signifie trop de paramètres, et donc un sur-apprentissage important (les performances du modèle sur les spectres d'apprentissage peuvent être bonnes, mais le modèle généralise mal).

La sélection de variables destinées à être utilisées dans un modèle non-linéaire reste une problématique importante. La complexité de ces modèles empêche une recherche exhaustive du meilleur ensemble de variables spectrales. Les méthodes linéaires classiques de sélection (comme l'analyse en composantes principales) ne conviennent pas non plus, car elles font perdre l'avantage potentiel d'une modélisation non-linéaire. Dans [BEN 04], une méthode de sélection de variables utilisant le modèle non-linéaire de prédiction a été présentée. La méthode sélectionne une variable à la fois, de façon à maximiser un critère de performance sur un ensemble de validation. Pour sélectionner une variable, il y a donc autant de modèles à tester qu'il reste de variables sélectionnables; la

[1] Michel Verleysen est Maître de Recherches du Fonds National de la Recherche Scientifique. Le travail de Damien François est financé par une bourse FRIA. Une partie de cet article présente des résultats de recherche financée par le programme belge des Pôles d'Attraction Interuniversitaires, mis en place par les Services fédéraux des affaires Scientifiques, Techniques et Culturelles de l'Etat belge. La responsabilité scientifique appartient à ses auteurs.

méthode, bien que performante, est dont extrêmement gourmande en ressources-calcul, ce qui la rend peu compatible avec la réalité de beaucoup d'applications. Par d'exemple, si des perceptrons multicouches (MLP) ou des réseaux à fonction radiales de base (RBFN) sont utilisés comme modèles non-linéaires de prédiction, des temps-calcul de plusieurs jours sur des machines performantes ne sont pas impossibles. Les performances de cette méthode ont été améliorées dans [BEN 05], en utilisant une mesure issue de la théorie de l'information, l'information mutuelle entre variables, pour sélectionner la première donnée spectrale. Malheureusement, la mesure utilisée dans [BEN 05] ne peut pas être étendue à la sélection des données spectrales suivantes; un retour à la méthode incrémentale précédente était alors indispensable, avec l' inconvénient majeur de la puissance-calcul nécessaire.

Cet article présente une nouvelle méthode de sélection de variables spectrales pour modèles non-linéaires. La méthode de sélection se base sur une extension du critère d'information mutuelle à un groupe de variables. Elle ne nécessite plus l'utilisation du modèle de prédiction lors du processus de sélection, diminuant ainsi de façon considérable le temps-calcul nécessaire. Comme elle se base uniquement sur l'information contenue dans les variables et non sur les modèles construits, elle peut être utilisée avec n'importe quel modèle (non-linéaire) de prédiction.

La section 2 de cet article décrit les détails de la méthode de sélection des variables spectrales grâce au critère d'information mutuelle, étendu à un groupe de variables. La section 3 montre les performances obtenues sur un problème traditionnellement utilisé comme "benchmark", la base de données Tecator. Les modèles non-linéaires utilisés sont les RBFN, ainsi que les Least-Square Support Vector Machines (LS-SVM). Les résultats sont comparés avec les méthodes traditionnelles, ainsi qu'avec les méthodes non-linéaires nécessitant des temps-calcul prohibitifs. La section 4 conclut sur l'utilité de la méthode proposée, comme compromis acceptable entre performances et temps-calcul compatible avec la réalité des applications.

2. Sélection des variables spectrales par information mutuelle

L'information mutuelle (*MI*) mesure la dépendance entre variables aléatoires; elle se différencie du coefficient de corrélation qui ne mesure que les dépendances linéaires entre ces variables. La *MI* est égale a zéro si et seulement si les variables sont strictement indépendantes et augmente avec la dépendance. Cette propriété est à la base de la méthode de sélection présentée ci-dessous.

Dans le cas de deux variables aléatoires X et Y, la *MI* est définie par :

$$MI(X,Y) = \iint \mu(X,Y) \log \frac{\mu(X,Y)}{\mu_x(X)\mu_y(Y)} dxdy ,$$

avec $\mu(X)$ et $\mu(Y)$ les distributions des deux variables, et $\mu(X,Y)$ leur distribution conjointe. La difficulté d'utiliser la *MI* pour la sélection de variables se situe dans l'estimation de ces densités de probabilité, par exemple par des méthodes à noyaux [SIL 86].

Une nouvelle méthode performante et rapide d'estimation de *MI* a récemment été publiée [KAR 04]. Cette méthode a pour premier avantage qu'elle ne nécessite pas le calcul des densités de probabilité, mais se base sur une estimation de l'entropie. En effet, la *MI* et l'entropie sont liées par la formule suivante :

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) ,$$

avec $H(.)$ l'entropie d'une ou plusieurs variables. Cette entropie est calculée par une méthode de plus proches voisins [KAR 04].

Un deuxième avantage important est que l'information mutuelle peut être calculée dans le cas où X et/ou Y sont des ensembles de plusieurs variables. La *MI* peut donc être utilisée pour calculer la dépendance entre un *ensemble* de variables d'entrées et une variable de sortie. Le meilleur jeu de variables d'entrées est celui qui maximise la *MI* et donc la dépendance entre un groupe de variables d'entrées et variable de sortie.

Malgré tout, dans un souci d'efficacité au point de vue temps-calcul, il n'est pas possible d'étudier la *MI* pour toutes les 2^n combinaisons de n variables d'entrée X_i. Si le calcul de ces 2^n combinaisons était possible dans des temps raisonnables, la combinaison

de variables donnant la *MI* la plus importante serait celle à retenir. Nous proposons donc une méthode itérative de sélection de variables selon le principe suivant.

1) L'information mutuelle entre chacune des variables d'entrées et la sortie est calculée. Les *n* variables sont classées par ordre décroissant selon cette information mutuelle. On a donc $MI(X_1,Y) > MI(X_2,Y) > \ldots > MI(X_n,Y)$.
2) L'ensemble des variables d'entrées X est initialisé avec la variable X_1. $X = [X_1]$. L'information mutuelle totale $MI(X, Y)$ entre X et Y est calculée.

Les étapes 3 à 4 sont répétées pour i allant de 2 à *n*.

3) On rajoute la variable X_i à l'ensemble des entrées X si cela provoque une augmentation de l'information mutuelle. Donc si $MI(X + [X_i], Y) > MI(X, Y)$ alors $X^{new} = X^{old} \cup [X_i]$.
4) Parmi tous les variables qui composent X, on recherche si la suppression d'une ou plusieurs de ces variables n'a pas pour conséquence une augmentation de l'information mutuelle totale. Si c'est le cas, cette ou ces variables sont éliminées de l'ensemble X. Donc si $MI(X \setminus [X_j], Y) > MI(X, Y)$ alors $X^{new} = X^{old} \setminus X_j$.

Dans l'étape 3, on rajoute les variables qui apporte une nouvelle information tandis que dans l'étape 4, on supprime les variables qui sont éventuellement devenues redondantes. L'importance de ces variables est mesurée par la *MI*. Il est important de maximiser l'information contenue dans les variables d'entrées tout en gardant un nombre d'entrées minimal, ceci afin d'éviter les problèmes de surapprentissage et ceux liés au grandes dimensions (*curse of dimensionality*). En général, les variables qui seront sélectionnées ne sont pas celles qui sont classées en premières dans l'étape 1. En effet, c'est l'information contenue dans un ensemble de variables qui est importante et non l'information contenue dans chacune des variables individuellement.

Cette méthode itérative, même si elle ne conduit pas à un ensemble optimal, a une complexité de l'ordre de n^2, plutôt que 2^n dans le cas d'une recherche exhaustive.

L'expérience montre (voir un exemple dans la section suivante) que l'ensemble de variables d'entrées ainsi sélectionné permet d'approcher de manière performante la sortie Y.

3. Résultats

Le benchmark classique Tecator [TEC] contient les spectres d'absorbance en proche infrarouge (850 à 1050 nm) d'échantillons de viande. Chaque spectre comporte 100 valeurs. Le but du benchmark est déterminer à partir du spectre le pourcentage de graisse contenu dans l'échantillon de viande correspondant. Pour ce faire, on dispose de 215 spectres qui ont été répartis par les producteurs de la base de données en 172 exemples d'apprentissage et 43 exemples de test. L'ensemble de test est uniquement utilisé pour évaluer les performances des algorithmes étudiés : on mesure celles-ci grâce à la NMSE (Normalized Mean Square Error). Cette approche permet d'éviter la surestimation des performances qui découlerait de l'estimation de celles-ci grâce à l'ensemble d'apprentissage.

Les algorithmes étudiés comportent des méta-paramètres (nombre de variables à retenir, nombre de neurones, etc.) dont les valeurs optimales doivent être déterminées par comparaison de performances. On estime les performances associées aux différentes valeurs des méta-paramètres par une validation croisée basée sur un découpage de l'ensemble d'apprentissage en 4 sous-ensembles contenant chacun 43 spectres.

Les spectres du benchmark Tecator comportent un effet moyen assez peu significatif: la valeur moyenne de chaque spectre est un très mauvais prédicteur du taux de graisses. Nous avons donc centré et réduit chaque spectre, puis ajouté deux variables additionnelles (la moyenne et l'écart-type des spectres d'origine) aux 100 variables initiales.

Les NMSE obtenues sur l'ensemble de test sont donnés dans le tableau 1. Sans surprise, on voit qu'un modèle linéaire sur des variables sélectionnées par une méthode

non-linéaire n'apporte rien, tandis que des modèles non-linéaires apportent une performance nettement améliorée. A titre de comparaison, des RBFN appliqués sur les coordonnées PCA donnent une erreur de 0.0121, avec l'inconvénient qu'il s'agit dans ce cas d'une méthode nécessitant un apprentissage de modèle à chaque étape de la sélection de variables, donc entraînant un temps-calcul de l'ordre de 20 fois supérieur. Un MLP sur ces mêmes variables permet de réduire l'erreur d'un ordre de grandeur, ce qui est considérable, mais au prix d'un temps-calcul environ 600 fois supérieur dans ce cas!

Méthode de prédiction	NSME (test)
PCR (régression en composantes principales)	0.0147
PLSR (régression en moindres carrés partiels)	0.0128
modèle linéaire sur variables sélectionnées par information mutuelle	0.0267
RBFN sur variables sélectionnées par information mutuelle	0.0064
LS-SVM sur variables sélectionnées par information mutuelle	0.0068

Tableau 1: résultats des différentes méthodes de prédiction sur la base de données Tecator (voir texte pour le détail des expériences). Les résultats sont exprimés en NMSE.

4. Conclusion

L'utilisation de modèles non-linéaires de prédiction pour des données de grandes dimensions comme des spectres nécessite la sélection d'un ensemble réduit de variables pertinentes. Pour ne pas perdre l'avantage des modèles non-linéaires de prédiction, la sélection de variables doit elle aussi être non-linéaire. Utiliser le modèle de prédiction dans la procédure de sélection entraîne des temps-calculs tout à fait prohibitifs par rapport à la réalité des applications. Cet article montre comment sélectionner un ensemble réduit de variables sans faire appel au modèle de prédiction, grâce à une mesure d'information mutuelle étendue aux ensembles de variables. Les performances de la méthode sont illustrées sur la base de données "Tecator"; il est montré que les performances restent acceptables par rapport aux techniques prohibitives en temps-calcul, tout en offrant des réelles possibilités de mise en œuvre dans des contextes applicatifs.

Références

[BEN 04] N. Benoudjit, E. Cools, M. Meurens, M. Verleysen, "Chemometric calibration of infrared spectrometers: Selection and validation of variables by non-linear models", Chemometrics and Intelligent Laboratory Systems, Elsevier, Vol. 70 (2004), No. 1, pp. 47-53.

[BEN 05] N. Benoudjit, D. Francois, M. Meurens, M. Verleysen, "Spectrophotometric variable selection by mutual information", accepté pour publication dans Chemometrics and Intelligent Laboratory Systems, Elsevier.

[BER 99] Bertran E., Blanco M., Maspoch S. et Pagès J., "Handling intrinsic non-linearity in near-infrared reflectance spectroscopy", Chemometrics and intelligent laboratory systems, 49: 215-224, 1999.

[BER 00] Bertrand D., Dufour E., "La spectroscopie infrarouge et ses applications analytiques", Eds Tec& Doc, collection sciences et techniques agroalimentaires, (2000).

[EKL 99] Eklov T, Martensson P., Lundstrom I, "Selection of variables for interpreting multivariate gas sensor data", Analytica Chimica Acta 381 (1999) 221-232.

[KAR 04] A. Kraskov, H. Stögbauer, P. Grassberger, “Estimating mutual information”, Phys. Rev. E, in press.

[MAS 97] Massart D. L., Vandeginste B. G. M., Buydens L. M. C., De Jong S., Lewi P. J., Smeyers-Verbeke J., "Handbook of Chemometrics and Qualimetrics : Part A", Elsevier Science, Amsterdam, 1997.

[SIL 86] B. W. Silverman, "Density Estimation". Chapman & Hall/CRC, London, 1986.

[TEC] Données spectrométriques Tecator, disponibles sur Statlib, http://lib.stat.cmu.edu/datasets/tecator