Chimiométrie 2003, Paris (France), 3-4 December 2003, pp. 133-136

Utilisation de l'information mutuelle pour la sélection de variables spectrales avec des modèles non-linéaires

Nabil **Benoudjit**[1], Damien **François**[2+], Marc **Meurens**[3], Michel **Verleysen**[1*]
Université catholique de Louvain (UCL),
[1]*Laboratoire de Microélectronique (DICE), 3 place du Levant, 1348 Louvain-la-Neuve (Belgique)*
{benoudjit, verleysen}@dice.ucl.ac.be
[2]*CESAME, 4 Avenue Georges Lemaître, 1348 Louvain-la-Neuve (Belgique)*
francois@auto.ucl.ac.be
[3]*Laboratoire de Spectrophotométrie (BNUT), 2(8) place Croix du Sud, 1348 Louvain-la-Neuve (Belgique)*
meurens@bnut.ucl.ac.be

MOTS-CLÉS: PCR (PRINCIPAL COMPONENT REGRESSION); PLSR (PARTIAL LEAST SQUARES REGRESSION); SMLR (STEPWISE MULTIPLE LINEAR REGRESSION); RBFN (RADIAL BASIS FUNCTIONS NETWORKS); FORWARD SELECTION; BACKWARD SELECTION ; MUTUAL INFORMATION

1. Introduction

La chimiométrie est [1] une branche de la chimie dans laquelle on utilise la théorie et les modèles développés en statistiques, mathématiques et informatique pour extraire l'information utile présente dans des données de mesures chimiques. Dans les applications analytiques, le cas le plus fréquent correspond à la prédiction d'une variable quantitative, telle que par exemple la concentration d'un composant présent dans un produit étudié.

D'un point de vue chimiométrique, les données spectrales ont des caractéristiques remarquables, qui nécessitent un traitement par des méthodes spécifiques [2]. La matrice X des données peut comporter plus de variables (données spectrales) que d'observations (spectres). Ainsi, certaines colonnes (variables) de la matrice X peuvent être représentées comme étant une combinaison linéaire d'autres colonnes de cette même matrice. Ce phénomène est connu sous le nom de 'colinéarité', et est la source de plusieurs problèmes dans l'application directe de la régression linéaire multiple [2] [3] [4]. Les études ont prouvés que si la colinéarité est présente parmi les variables, les résultats de la prédiction seront médiocres [3]. C'est pourquoi plusieurs méthodes linéaires ont été proposées pour éviter les problèmes liés à la redondance entre les variables, telles que la régression linéaire multiple pas-à-pas (SMLR) [2][5], la régression en composantes principales (PCR) [2][4][6] et la régression des moindres carrés partiels (PLSR) [2][4].

Bien souvent les techniques de calibrage multivarié des spectrophotomètres infrarouges pour l'analyse chimique quantitative font l'hypothèse de l'existence d'une relation linéaire entre les variables dépendantes (concentration des analytes) et les variables indépendantes (données spectrales), alors qu'une véritable relation strictement linéaire n'existe que très rarement. Les méthodes linéaires PCR et PLSR basées sur la condensation des données spectrales fournissent des résultats acceptables aussi longtemps que la déviation par rapport aux linéarités n'est pas trop grande. Cependant, en présence de non-linéarités fortement prononcées, les deux techniques de calibrage peuvent mener à des erreurs importantes. D'autres méthodes chimiométriques doivent alors être utilisées [7]. La non-linéarité en analyse infrarouge résulte de divers facteurs, à savoir les déviations par rapport à la loi de Beer-Lambert, la réponse non-linéaire des capteurs, les dérives dans la source lumineuse, etc.[7][8].

Nous avons proposé précédemment [9-10] une procédure de sélection de données spectrales en moyen infrarouge et en proche infrarouge basée sur la combinaison de trois mécanismes (régression non-linéaire RBFN, procédure incrémentale de sélection des variables et utilisation d'un ensemble de validation),

* Michel Verleysen est Maître de Recherches du Fonds National de la Recherche Scientifique belge.
+ Le travail de D. François est financé 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 auteur(s).

améliorant les performances de prédiction par rapport aux méthodes linéaires de sélection. Nous proposons dans ce travail d'améliorer cette méthode par un choix judicieux de la première donnée spectrale, qui a une très grande influence sur les performances finales de prédiction. L'idée est d'utiliser une mesure de l'information mutuelle entre les données spectrales (variables indépendantes x_i) et la concentration d'analyte (variable dépendante y) pour sélectionner la première variable; la procédure incrémentale [9-10] est alors utilisée pour sélectionner les variables spectrales suivantes. Si l'idée d'utiliser un critère d'information mutuelle pour le choix des autres données spectrales peut sembler également pertinente, sa mise en œuvre devient de plus en plus difficile lorsque le nombre de variables sélectionnées augmente (phénomène d'espace vide [11]). Nous limiterons donc l'utilisation du critère d'information mutuelle au choix crucial de la première variable spectrale, et nous verrons dans la section résultats qu'il est avantageux de combiner ainsi les deux approches.

2. Information mutuelle

Dans cette section nous allons voir comment l'information mutuelle peut être utilisée pour évaluer l'importance de chacune des variables (données spectrales) d'entrée des modèles à calibrer.

2.1. Définition

Le but principal d'un modèle de prédiction est de réduire au maximum l'incertitude sur la variable de sortie. Un bon formalisme de l'incertitude d'une variable aléatoire est donné par la théorie de l'information de Shannon [12]. Initialement développé d'abord pour des variables binaires, elle peut être étendue aux variables continues. L'incertitude d'une variable aléatoire *Y* à valeurs *y* dans l'ensemble fini *D* peut être mesurée au moyen de son entropie *H*:

$$H(Y) = -\sum_{D} P(Y = y) \cdot \log P(Y = y).$$

Pour illustrer cette notion, supposons que dans un cas extrême toutes les valeurs dans *D* aient une probabilité nulle excepté une, qui a la probabilité égale à 1. Alors, il n'y a absolument aucune incertitude puisque *Y* est une constante, et *H*(*Y*)=0. Supposons maintenant que toutes les valeurs dans *D* soient équiprobables. L'incertitude est alors totale et $H(Y) = \log N$, que l'on peut démontrer être le maximum de cette fonction. Lorsque la valeur d'une autre variable *Z* à valeurs dans *D'* est connue, on peut définir l'entropie conditionnelle :

$$H(Y \mid Z) = -\sum_{D'} P(Z = z) \sum_{D} P(Y = y \mid Z = z) \cdot \log P(Y = y \mid Z = z).$$

L'information mutuelle entre *Z* et *Y* est alors définie par :

$$I(Y, Z) = H(Y) - H(Y \mid Z),$$

le dernier terme représentant la diminution d'incertitude sur *Y* quand *Z* est connu.

Les notions d'entropie, entropie conditionnelle et information mutuelle, peuvent être étendues au cas continu (ensemble *D* de taille infinie). L'information mutuelle entre les variables *Y* et *Z* devient par exemple

$$I = \int h(z, y) \cdot \log \frac{h(z, y)}{f(z) \cdot g(y)} dz dy ,$$

où f(z) et g(y) sont les densités de probabilités marginales des variables *Z* et *Y*, et h(z,y) est la fonction de densité de probabilité jointe de *Z* et de *Y*. Cette formulation montre que l'information mutuelle entre *Z* et *Y* vaut zéro si et seulement si *Z* et *Y* sont statistiquement indépendant. Nous pouvons remarquer que l'information mutuelle n'est affectée par aucune transformation de variable, et ne fait aucune hypothèse sur la relation entre *Z* et *Y*.

2.2. Calcul de l'information mutuelle

Le calcul de l'information mutuelle est basé sur l'estimation des fonctions de densité de probabilités et probabilités jointes (pdf) des variables. Cette estimation doit être effectuée sur base de données, en utilisant généralement soit des histogrammes soit des noyaux [13]. Dans ce travail nous utilisons les histogrammes pour leurs avantages indéniables en termes de complexité de calcul, et leurs performances suffisantes dans ce contexte. L'estimation des densités de probabilités jointes passe par l'utilisation d'histogrammes bi-dimensionnels. Les tailles des cellules (bins) des histogrammes sont des paramètres importants qui doivent être choisis soigneusement. Si les cellules sont trop grandes, l'approximation ne sera pas assez précise ; si elles sont trop petites, la plupart d'entre elles seront vides et l'approximation ne sera pas suffisamment lisse. Même si des heuristiques ont été proposées [14][15][16] pour guider ce choix, seule l'expérience peut conduire à un choix optimal. Dans notre cas nous nous limiterons aux grilles régulières, dont les cellules sont de tailles identiques, et nous choisirons la taille des cellules selon un procédé de validation.

3. Sélection et validation des variables par modèles non-linéaires

La procédure de sélection des variables qui seront utilisées dans le modèle non-linéaire de prédiction, parmi les p variables spectrales disponibles, est alors la suivante :

1. la première donnée spectrale sélectionnée est celle qui maximise l'information mutuelle avec la variable dépendante, selon le calcul détaillé dans la section 2 ;
2. les autres données spectrales sont sélectionnées selon la procédure « Forward-Backward selection » [9-10] :
   - Forward selection : une fois que k variables sont sélectionnées (k=1 à la première itération), p-k modèles à k+1 variables (les k sélectionnées et respectivement chacune des autres) sont construits et comparés par un critère NMSE (Normalised Mean Square Error) ; la variable correspondant au minimum du critère est rajoutée. Le processus "forward" est répété pour k=2,3,…, jusqu'à ce que la valeur du critère NMSE (sur un ensemble indépendant de validation) augmente.
   - Backward selection : consiste à éliminer les données spectrales les moins significatives sélectionnées dans l'étape "forward". Si q variables spectrales ont été sélectionnées lors de l'étape "forward", q modèles sont construits en enlevant respectivement chacune des q variables, et leurs NMSE sont calculés. Celui qui minimise le critère est sélectionné. Une fois le modèle choisi, son NMSE avec celui du modèle obtenu à l'étape précédente. Si le nouveau NMSE est inférieur à celui de l'étape précédente, alors la donnée spectrale éliminée est non-significative (redondante). Le processus est alors répété pour les données spectrales restantes. Dans le cas contraire, la donnée spectrale choisie pour être éliminée est significative, et le processus de 'backward selection' est arrêté.

Il faut noter que lors de l'étape 1, nous n'avons besoin que de l'ensemble d'apprentissage, car le calcul de l'information mutuelle ne nécessite pas l'estimation et la comparaison de modèles. Par contre pour l'étape 2, l'utilisation d'autres données (ensemble de validation), indépendantes de l'ensemble d'apprentissage, pour le calcul des NMSE est indispensable pour détecter et éviter le phénomène de sur-apprentissage (overfitting) [10].

4. Résultats

La base de données utilisée comprend les spectres (700 données d'absorbance proche infrarouge) et les concentrations en sucre (saccharose) de 218 échantillons de jus. 150 spectres ont été utilisés pour l'apprentissage, 68 spectres ont été utilisés pour la validation du choix des variables.

Méthodes de calibration	Nombre de variables	$NMSE_V$
PCR	42	0.2596
PLSR	16	0.2435
SMLR	16	0.5137
FBS-RFBN	13	0.0691
IM+FBS-RBFN	**36**	**0.0313**

Table 1 : Erreurs quadratiques moyennes normalisées obtenues sur un ensemble de validation avec les 5 méthodes.

Cette table nécessite les commentaires suivants :

- Le modèle 'FBS-RBFN' correspond à la procédure décrite dans [9][10]. Le réseau RBFN utilisé est constitué d'une seule couche cachée avec 8 fonctions gaussiennes; sa procédure d'apprentissage est décrite dans [17].
- Le modèle 'IM+FBS-RBFN' correspond à la procédure décrite dans ce papier. Le nombre optimal de fonctions gaussiennes utilisé dans le réseau RBFN est 5.
- Nous remarquons que l'erreur de prédiction avec la méthode IM+FBS-RBFN est deux fois plus petite que l'erreur obtenue avec la méthode FBS-RBFN. La première donnée spectrale sélectionnée par chacune des deux méthodes est différente (x_{256} dans le premier cas et x_{690} dans le second).

5. Conclusion

Nous avons proposé une amélioration de la procédure de sélection de données spectrales basée sur la combinaison de trois mécanismes (régression non-linéaire, procédure incrémentale de sélection des variables et utilisation d'un ensemble de validation). Cette amélioration a permis d'une part de profiter du potentiel des méthodes non-linéaires pour prédire une donnée chimique qui n'est probablement pas en relation tout à fait linéaire avec le spectre proche infrarouge du produit analysé, et d'autre part d'éviter le sur-apprentissage des données. Les résultats obtenus avec la nouvelle approche montrent clairement que les performances sont bien meilleures; ceci montre l'avantage de notre nouvelle approche du problème.

6. Références

[1] Geladi P. and Dabakk E., An overview of chemometrics application in NIR spectrometry, J. NIR Spectrosc., 3: 119-132, 1995.

[2] Bertrand D., Dufour E., *La spectroscopie infrarouge et ses applications analytiques*, Editions Tec& Doc, collection sciences et techniques agroalimentaires, (2000).

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

[4] Geladi P., Kowalski B. R., *Partial least squares regression : A Tutorial*, Analytica Chimica Acta, 185 (1986) 1-17.

[5] 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.

[6] A. D. Walmsley, *Improved variable selection procedure for multivariate linear regression*, Analytica Chimica Acta, 354 (1997) 225-232.

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

[8] Blanco M., Coello J., Iturriaga H., Maspoch S. and Pagès J., *Calibration in non-linear near infrared reflectance spectroscopy : a comparison of several methods*, Analytica Chimica Acta, 384: 207-214, 1999.

[9] N. Benoudjit, E. Cools, M. Meurens and M. Verleysen, *Calibrage chimiométrique des spectrophotometers: selection et validation des variables par modèles non-linéaires*, Proccedings Chimiométrie 2002, Paris (France), 4-5 Decembre 2002, pp. 25-28.

[10] N. Benoudjit, E. Cools, M. Meurens and M. Verleysen, *Chemometric calibration of infrared spectrometers: Selection and validation of variables by non-linear models*, Accepted for publication in Chemometrics and intelligent laboratory systems, Elsevier.

[11] Scott, D. W., Thompson, J. R., *Probability density estimation in higher dimensions*, In Douglas, S. R. (ed): Computer Science and Statistics, Proceedings of the Fifteenth Symposium on the Interface, North Holland-Elsevier, Amsterdam, New York, Oxford (1983) pp. 173-179.

[12] C. E. Shannon and W. Weaver, The Mathematical Theory of Communication, Urbana, IL: University of Illinois Press, 1949.

[13] D. W. Scott, Multivariable Density Estimation: Theory, Practice, and Visualization. John Wiley, New York, 1992.

[14] B. V. Bonnlander and A. S. Weigend, Selecting input variables using mutual information and nonparametric density estimation. In Proceedings of International Symposium on Artificial Neural Networks (ISANN'94), 1994.

[15] A. J. Izenman, Recent developments in nonparametric density estimation, Journal of the American Statistical Association, 86(413): 205-224, 1991.

[16] D. Scott, On optimal and data-based histograms, Biometrika, 66: 605-610, 1979.

[17] Benoudjit N., Archambeau C., Lendasse A., Lee J., Verleysen M., *Width optimization of the Gaussian kernels in Radial Basis Function Networks*, ESANN (2002), April 24-25-26, p. 425-432, Bruges.