Chimiométrie 2002 proceedings
Paris (France), 4-5 December 2002, pp. 25-28

Calibrage chimiométrique des spectrophotomètres : sélection et validation des variables par modèles non-linéaires

Nabil Benoudjit[1], Etienne Cools[2], Marc Meurens[2], Michel Verleysen[1*]

Université catholique de Louvain,
[1]Laboratoire de Microélectronique (DICE), 3 place du Levant, 1348 Louvain-la-Neuve (Belgique)
{benoudjit, verleysen}@dice.ucl.ac.be
[2]Laboratoire de Spectrophotométrie (BNUT), 2(8) place Croix du Sud, 1348 Louvain-la-Neuve (Belgique)

Résumé. Les données acquises par les spectrophotomètres constituent des spectres. Il s'agit d'ensemble d'un grand nombre de variables exploitables en analyse chimique quantitative moyennant l'établissement de modèles de calibrage par des méthodes chimiométriques. Pour établir ces modèles de calibrage qui sont spécifiques à chaque paramètre analysé, il convient de sélectionner un nombre réduit de variables spectrales. Ce papier présente une nouvelle méthode incrémentale (pas-à-pas) de sélection des variables spectrales par calculs de régression linéaire et de réseau neuronal, basée sur une validation objective (externe) du modèle de calibrage ; cette validation est effectuée sur des ensembles indépendants de données correspondant à d'autres échantillons (des mêmes produits) que ceux utilisés lors du calibrage. Les avantages de la méthode présentée sont discutés et mis en évidence par rapport aux méthodes de calibrage actuellement utilisées en analyse chimique quantitative par spectrophotométrie.

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

1. Calibrage chimiométrique des spectrophotomètres

L'analyse chimique par spectrophotométrie repose sur l'acquisition rapide d'un grand nombre de données spectrales (plusieurs centaines, voire plusieurs milliers). Lorsque ces données ne sont pas condensées dans des vecteurs propres par l'analyse en composantes principales, seul un petit nombre d'entre elles peuvent entrer dans la constitution de chaque modèle de calibrage pour la détermination d'un constituant particulier. Les modèles de calibrage consistent véritablement en des équations de conversion des données spectrales (entrée) en des valeurs de composition chimique (sortie). L'établissement de ces équations suppose l'ajustement des paramètres affectant les valeurs d'entrée pour arriver aux valeurs de sortie les plus proches possibles de la réalité.

L'inutilité de certaines données spectrales, ainsi que la difficulté d'ajuster des modèles comportant un trop grand nombre d'entrées rend, leur sélection obligatoire.

[*] Michel Verleysen est Maître de Recherches du Fonds National de la Recherche Scientifique belge.

Chimiométrie 2002 proceedings
Paris (France), 4-5 December 2002, pp. 25-28

Nous présenterons d'abord les techniques habituelles de sélection de variables correspondant aux méthodes de régression linéaire : ce sont les méthodes de calibrage dites SMLR (régression linéaire multiple pas-à-pas), PCR (régression des composantes principales) et PLSR (régression des moindres carrés partiels). Ensuite, nous proposerons d'incorporer des modèles de régression non-linéaires (RBFN – Réseaux de neurones à Fonctions Radiales de Base) pour la sélection de variables, au travers d'une procédure incrémentale basées sur un critère de validation. Enfin, nous présenterons une comparaison de résultats de prédiction de la teneur en alcool obtenus avec les différents modèles de calibrages sur des spectres infrarouges (FTIR) de vin.

2. Sélection des variables : état de l'art

Les méthodes de sélection des variables spectrales par calcul de régression linéaire habituellement utilisées pour calibrer les spectrophotomètres sont les suivantes :

- SMLR (régression linéaire multiple pas-à-pas) : les données spectrales sont sélectionnées parmi les *p* données disponibles en respectant un critère d'optimisation tel que le test d'hypothèse basé sur la loi de Fisher ; celui-ci permet de juger le caractère significatif de l'ajout ou de la suppression d'une variable. Nous appliquons en général une succession d'étapes ascendantes (forward), dans lesquelles une donnée spectrale est introduite à chaque étape, suivies d'étapes descendantes (backward), dans lesquelles la variable la moins pertinente est éliminée à chaque étape [BER 00][MAS 97].

- PCR (régression en composantes principales) : consiste à appliquer tout d'abord une analyse en composantes principales (ACP) sur la matrice des données spectrales. L'ACP permet de remplacer les données spectrales d'origine, fortement redondantes, par des composantes principales (combinaisons linéaires des données d'origine), qui contiennent la quasi-totalité de l'information, et qui ont l'avantage d'être non corrélées, ou orthogonales entre elles. Les données spectrales condensées par l'ACP peuvent servir alors de variables de base à une régression linéaire multiple [BER 00][GEL 86][WAL 97].

- PLSR (régression des moindres carrés partiels): consiste en une régression de la caractéristique à prédire sur des variables latentes (combinaisons linéaires des données spectrales). Dans cette méthode, les variables latentes sont déterminées en tenant compte de la sortie (caractéristique à prédire) désirée du modèle et des données spectrales, alors que dans la PCR, elles sont déterminées sans tenir compte de la sortie désirée du modèle [BER 00][GEL 86].

Il faut noter que, contrairement à certains travaux publiés, dans tous les cas, le modèle devrait être vérifié sur un autre ensemble d'échantillons que ceux qui ont servi au calibrage proprement dit.

Tous ces modèles font l'hypothèse de l'existence d'une relation linéaire entre les variables sélectionnées ou construites d'une part, et la caractéristique à prédire d'autre part. Ceci peut évidemment ne pas être le cas dans la réalité de certaines applications. Certains auteurs utilisent des modèles non-linéaires, mais après une projection de type ACP ne tenant donc pas compte des valeurs désirées de la caractéristique à prédire [EKL 99]. Enfin, les méthodes incrémentales de type SMLR font appel à un critère de sélection calculé sur un ensemble d'apprentissage, et non de validation.

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

Au vu de ces limitations, nous proposons une méthode de sélection de variables basée sur les principes suivants :

- utilisation d'un modèle ***non-linéaire*** de régression (RBFN) ;
- choix des variables basé sur une procédure ***incrémentale*** (forward-backward);
- variables choisies en fonction du MSE (erreur quadratique moyenne) sur un ensemble de ***validation***.

Chimiométrie 2002 proceedings
Paris (France), 4-5 December 2002, pp. 25-28

Les RBFN [HAY 99] sont des modèles de régression non-linéaire ayant la propriété d'approximation universelle. Ils se basent sur une combinaison linéaire de fonctions gaussiennes, dont les centres et largeurs sont des paramètres supplémentaires. Pour des performances similaires, les RBFN offrent souvent un apprentissage plus aisé que les plus traditionnels réseaux MLP (perceptrons multi-couches).

Nous proposons une méthode de sélection de données spectrales basée sur un critère de validation MSE dite 'forward-backward selection'. La sélection des données spectrales est divisée en deux étapes.

La première étape est la 'forward selection'. Elle commence par la construction des p modèles possibles à une variable spectrale seulement. Nous calculons le critère MSE pour chacun de ces modèles et nous choisissons celui qui minimise le critère. Nous fixons ensuite la donnée spectrale déjà sélectionnée ; p-1 modèles sont alors construits en ajoutant une seule des variables spectrales restantes. Le critère MSE pour chacun de ces modèles est calculé, et nous choisissons le modèle qui minimise ce critère. Nous continuons le processus ci-dessus jusqu'à ce que la valeur du critère MSE augmente.

La deuxième étape est la 'backward selection'. Elle consiste à éliminer les données spectrales les moins significatives déjà sélectionnées dans la première étape. Si q variables spectrales ont été sélectionnées lors de la première étape, q modèles sont construits en enlevant une des variables sélectionnées. Le critère MSE est calculé sur chacun de ces modèles, et celui qui minimise le critère est sélectionné. Une fois le modèle choisi, nous comparons son critère MSE avec celui du modèle obtenu à l'étape précédente. Si le nouveau MSE est inférieur à celui de l'étape précédente, alors la donnée spectrale éliminée est non-significative. 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.

Habituellement, les procédures incrémentales de type SMLR utilisent un critère de sélection de type Fisher ou coefficient de détermination (R^2) de la régression. L'utilisation du test de Fisher permet de juger la pertinence d'une variable en comparant sa valeur de test à un seuil contenu dans une table. Néanmoins, les tables de Fisher ne sont valables que dans le cas où le critère est évalué sur les mêmes données que celles qui ont permis l'apprentissage du modèle. L'utilisation d'autre données (ensemble de validation) est néanmoins indispensable pour détecter et éviter le phénomène de sur-apprentissage (overfitting). La solution consiste donc à conditionner le choix d'une variable à une mesure des performances du modèle incluant cette variable sur un ensemble de validation ; dans notre cas, nous utiliserons le critère MSE.

La combinaison des trois concepts sous-jacents (régression non-linéaire, procédure incrémentale et choix basé sur un ensemble de validation) permet d'une part de profiter du potentiel des méthodes non-linéaires pour prédire un phénomène physique qui n'est probablement pas lui-même linéaire, et d'autre part d'éviter le sur-apprentissage des données. Cette procédure de sélection de variables offre donc, potentiellement, de meilleures performances lorsque celles-ci sont mesurées sur des données indépendantes de l'apprentissage. Ceci sera illustré par un exemple dans la section qui suit.

4. Résultats

La base de données utilisée comprend les spectres (256 données d'absorbance moyen infrarouge) et les teneurs en alcool de 124 échantillons de vin. 94 spectres ont été utilisés pour l'apprentissage, 30 spectres ont été utilisés pour la validation du choix des variables. Des expériences similaires ont été effectuées sur les spectres bruts et sur les mêmes spectres centrés et réduits.

Méthodes	PCR	PLSR	SMLR	Forward-Backward (Linéaire)	Forward-Backward (Non-linéaire)
# variables	30	12	14	17	20
MSE	0.0061	0.0106	0.0080	0.0024	**0.0019**

Table 1 : Erreurs quadratiques moyennes (MSE) obtenues sur un ensemble de validation avec les 5 méthodes en utilisant les spectres bruts.

Méthodes	PCR	PLSR	SMLR	Forward-Backward (Linéaire)	Forward-Backward (Non-linéaire)
# variables	20	11	14	15	23
MSE	0.0217	0.0238	0.0401	0.0044	**0.0033**

Table 2 : Erreurs quadratiques moyennes (MSE) obtenues sur un ensemble de validation avec les 5 méthodes en utilisant les spectres centrés et réduits.

Ces deux tables nécessitent les commentaires suivants :

- Malgré des erreurs quadratiques moyennes (MSE) très bonnes (0.0022 et 0.0009 respectivement pour les spectres bruts et ceux centrés et réduits) obtenues par la méthode SMLR sur l'ensemble d'apprentissage, les deux tables 1 et 2 montrent que le même critère calculé sur un ensemble de validation donne des erreurs quatre fois plus grandes et même plus, ce qui illustre bien la nécessité de travailler avec un ensemble de validation et non d'apprentissage.

- Le modèle 'forward-backward' linéaire correspond à la procédure décrite dans la section précédente, mis à part le fait que le modèle de régression utilisé est un modèle linéaire (et non un RBFN).

- Le modèle 'forward-backward' non-linéaire correspond à la procédure décrite dans la section précédente. Le réseau RBF utilisé dans les deux dernières expèriences des tables 1 et 2 est constitué d'une seule couche cachée avec 3 fonctions gaussiennes ; sa procédure d'apprentissage est décrite dans [BEN 02].

5. Conclusion

Nous avons proposé une procédure de sélection de données spectrales basée sur la combinaison des trois «mécanismes» (régression non-linéaire, procédure incrémentale de sélection des variables et utilisation d'un ensemble de validation). Cette procédure permet 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 infrarouge du produit analysé, et d'autre part d'éviter le sur-apprentissage des données. Les modèles non-linéaires, couplés à des procédures justifiées de sélection de variables, devraient permettre à l'avenir d'améliorer les performances de calibrage des spectrophotomètres ; les résultats obtenus montrent l'avantage de notre approche du problème.

6. Références

[BEN 02] 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.

[BER 00] Bertrand D., Dufour E., *La spectroscopie infrarouge et ses applications analytiques*, Editions 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.

[GEL 86] Geladi P., Kowalski B. R., *Partial least squares regression : A Tutorial*, Analytica Chimica Acta, 185 (1986) 1-17.

[HAY 99] Haykin S., *Neural Networks a Comprehensive Foundation*, Prentice-Hall Inc, second edition, 1999.

[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.

[WAL 97] A. D. Walmsley, *Improved variable selection procedure for multivariate linear regression*, Analytica Chimica Acta, 354 (1997) 225-232.