Chimiométrie 2005
Villeneuve d'Ascq (France), 30 November - 1 December 2005, pp. 143-146.

Estimation non paramétrique de bruit pour la construction de modèles non linéaires en spectrométrie

Amaury **Lendasse**[1], Damien **François**[2], Vincent **Wertz**[2], Michel **Verleysen**[3*]
[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
[2-3] *Université catholique de Louvain – Machine Learning Group,*
[2] *CESAME, 4 av. G. Lemaître, 1348 Louvain-la-Neuve, Belgique,* francois@auto.ucl.ac.be
[3] *DICE, 3 place du Levant, 1348 Louvain-la-Neuve, Belgique,* verleysen@dice.ucl.ac.be

MOTS CLÉS : Spectrométrie, modèles non linéaires multivariés, Estimation non paramétrique de bruit, PLS

1. Introduction

Dans le domaine de la spectroscopie analytique, le développement de modèles de prédiction précis et adaptés au problème étudié est un besoin incontestable. Dans de nombreuses applications, 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 utilisés [MAS 97]. Les avantages de tels modèles résident dans le fait qu'ils sont éprouvés et utilisés à bon escient par les praticiens, ainsi que dans la facilité de leur mise en œuvre (peu de paramètres à choisir par l'utilisateur, etc.). Dans certaines applications 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.

Lorsqu'un modèle est construit, il est souvent difficile de différencier le bruit, dû aux données, de l'erreur de modélisation. Lorsqu'un modèle non-linéaire est construit, il convient de l'ajuster pour diminuer le plus possible l'erreur de modélisation ; cet ajustement peut se faire par le choix d'une complexité adéquate de modèle (nombre de paramètres), ou d'une sélection appropriée de ses entrées lorsque le nombre de variables potentielles est grand (ce qui est le cas avec des spectres). Néanmoins, il convient de ne pas trop augmenter la complexité et/ou le nombre d'entrées au modèle ; en cas d'excès, le sur-apprentissage (overfitting) apparaît : le modèle convient bien aux données d'apprentissage, mais devient inutilisable car il généralise mal sur de nouvelles données.

Comme l'erreur observée sur un modèle est la somme de l'erreur de modélisation et du bruit sur les données, et que ce dernier est constant quelque soit le modèle utilisé, rendre minimum l'erreur observée est équivalent à rendre minimum l'erreur de modélisation. Néanmoins il est en général impossible, autant pour des raisons de colinéarité entre variables spectrales que de temps-calcul, de tester un grand nombre de modèles non-linéaires différents. Il est donc difficile, voir impossible en pratique, de sélectionner les variables spectrales les plus adéquates à utiliser comme entrée au modèle.

Pour remédier à cette difficulté, cet article présente une méthode d'estimation non-paramétrique de bruit [JON 04]. L'avantage de cette technique est qu'elle ne nécessite pas l'apprentissage coûteux d'un modèle non-linéaire pour évaluer le bruit ; elle peut donc être utilisée de manière répétitive pour des ensembles différents (et de taille différente) de variables spectrales, afin d'évaluer lesquelles sont les plus adéquates à être utilisée dans un modèle (non-linéaire) tout en évitant le sur-apprentissage. Cette méthode permet également la sélection des meilleurs paramètres de pondération des entrées, ce qui apparaît

[*] 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.

Chimiométrie 2005
Villeneuve d'Ascq (France), 30 November - 1 December 2005, pp. 143-146.
souvent comme un problème crucial en modélisation non-linéaire. Les temps de calculs nécessaires à la méthode restent incompatibles avec les données spectrométriques (trop grand nombre de variables) ; la méthode est donc associée à la régression PLS afin de ne travailler que sur un nombre réduit de variables, les variables latentes de la régression PLS. Les méthodes proposées (PLS + sélection et PLS + pondération) sont des extensions non linéaires de la régression PLS classique ; elles combinent les avantages de la régression PSL (temps de calcul) et des approximateurs non linaires (précisions). La section 2 décrit brièvement l'estimation paramétrique du bruit, les méthodes de sélection et de pondération de variables, ainsi que leur utilisation dans le cadre de la régression PLS. 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 Least-Square Support Vector Machines (LS-SVM, [SUY 02]).

2. Estimation non-paramétrique de bruit

2.1 Delta Test

L'estimation non-paramétrique de bruit a été développée dans le cadre de la sélection d'entrées pour des problèmes d'approximation de fonction. Le problème est défini par I paires entrées-sortie $(\mathbf{x}^i, y^i)$ qui appartiennent à R^MxR. La relation entre $\mathbf{x}_i$ et y_i peut s'exprimer comme

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \qquad (\mathbf{1})$$

où f est la relation inconnue et ε_i le bruit. Une approche appelée "Delta Test" a été proposée afin d'estimer la variance du bruit du modèle (1) [JON 04]. L'approche se base sur la similarité entre deux voisins dans l'espace des entrées : si la distance δ entre 2 points $\mathbf{x}$ et $\mathbf{x}$' tend vers zéro, la distance moyenne (divisée par deux) entre les deux sorties correspondante tend vers la variance du bruit [JON 04]:

$$E\left\langle \frac{1}{2}(y'-y)^2 \;\middle|\; |\mathbf{x}'-\mathbf{x}| < \delta \right\rangle \to \mathrm{var}(\varepsilon) \quad \text{si} \quad \delta \to 0 \qquad (\mathbf{2})$$

Les hypothèses sous-jacentes à cette approche ont les suivantes :

- Les dérivées première et seconde du modèle f doivent être bornées.
- Le bruit doit être de moyenne nulle. Si ce n'était pas le cas, cette moyenne peut être incorporée dans le modèle f pour que l'hypothèse soit respectée.

Ces hypothèses sont des hypothèses classiques dans le cadre de l'approximation de fonctions et cadrent avec les données spectrométriques. Il est donc possible d'estimer relativement précisément la variance du bruit sur un modèle. Cette estimation sera de plus en plus précise si le nombre de données est grand. S'il est faible, des améliorations du Delta Test ont été proposées [JON 04].

L'estimation ainsi calculée dépend des M entrées qui ont été utilisées. Si ces entrées sont pertinentes la variance du bruit sera faible et par conséquent $f(\mathbf{x})$ sera une bonne approximation de y. Inversement, si les entrées ne sont pas pertinentes la variance du bruit sera élevée et $f(\mathbf{x})$ ne sera pas une bonne approximation de y. L'estimation de la variance est donc un bon critère pour sélectionner les entrées pertinentes, dans ce cas les variables spectrales. Le meilleur ensemble d'entrées est celui qui minimisera la variance du bruit.

Le point faible de la méthode reste que si le nombre de variables spectrales disponibles est N (desquelles M seront sélectionnées), il existe 2^N-1 ensemble d'entrées possibles. Pour N élevé, il n'est donc pas possible d'évaluer l'estimation de la variance pour chacune des 2^N-1 possibilités. Afin, de palier à cette faiblesse, deux nouvelles méthodes combinant régression PSL et Delta Test sont proposées dans la section suivante.

2.1 Delta Test et PLS pour la sélection de variables

La régression PLS [MAS 97] est une méthode de régression linéaire utilisant un nombre restreint P de variables latentes résultant d'un compromis : elle doivent représenter au mieux les N variables initiales et fournir une bonne approximation de la sortie désirée.

Chimiométrie 2005
Villeneuve d'Ascq (France), 30 November - 1 December 2005, pp. 143-146.

Les variables initiales sont **x** sont décomposées linéairement en variables latentes **t**: $x = t.p' + e$ où **e** représente l'erreur de décomposition et **p** le poids de chacune des variables latentes dans la décomposition. La sortie y est approximée par le modèle linéaire suivant : $y = t.q' + h$ où q est le vecteur des coefficients de la régression linéaire et h l'erreur d'approximation.

Si le nombre de variables latentes augmente, l'erreur de décomposition tendra vers zéro. En vue d'obtenir une bonne généralisation du modèle linéaire, le nombre de variables latentes doit être optimisé : celui-ci doit être suffisamment élevé afin d'approximer au mieux la sortie y, pas trop afin d'éviter le sur apprentissage. Les méthodes itératives permettant de calculer t, p et q sont présentées dans [MAS 97]. Il est important de noter que l'importance des variables latentes est décroissante car celles-ci résultent de la décomposition en valeurs singulières du produit matriciel **x**.y [MAS 97]. Les dernières variables latentes ne représentent donc plus que du bruit. Le nombre de variable latente peut être déterminé en utilisant le Delta Test qui a été présenté dans la section précédente. Il est calculé successivement pour un nombre croissant de variables latentes ; le nombre optimal de variables sera celui qui minimisera l'estimation de la variance du bruit. Le nombre d'estimation de la variance du bruit est donc fortement réduit : de 2^N à $2N$ estimations ; cette réduction est due à l'ordonnancement effectué par la PLS.

2.2 Delta Test et PLS pour la pondération de variable

Une nouvelle application du Delta Test est la détermination automatique de la pondération des variables sélectionnées. En effet, certains approximateurs non-linéaires (et en particulier ceux basés sur des noyaux) sont très sensibles à la pondération relative des différentes variables ; il est donc important de pouvoir différencier cette importance entre les variables. La sélection des variables est un cas particulier de pondération où les poids sont égaux à zéro ou à 1. Le Delta Test permet de déterminer si une pondération particulière des variables d'entrées donne une estimation plus faible de la variance et donc un meilleur modèle. Le Delta Test devant être effectué une fois la pondération choisie, il est nécessaire de discrétiser la pondération de chaque variable en k niveaux. Le nombre de tests à effectuer sur les variables spectrales initiales (k^M) resterait néanmoins prohibitif (la sélection de variables est un cas particulier avec k=2). Une fois encore, l'ordonnancement des variables latentes effectué par la PLS permet de résoudre ce problème. Le poids de la première variable latente est choisi égal à 1 sans perte de généralité. Par la suite, le poids de la deuxième variable latente est choisi dans l'ensemble $\{0, 1/k, 2/k, \ldots, 1\}$. Le poids de la troisième variable latente est choisi de façon similaire, et ainsi de suite. Quand l'ajout de nouvelles variables latentes devient inutile, tout les poids optimaux deviennent égaux à 0. En général, k=10 est une discrétisation suffisante des poids.

Cette méthode est très rapide car elle ne nécessite que $k.N$ estimation de la variance du bruit. La pondération ainsi obtenue n'est pas optimale mais ne peut que donner de meilleurs résultats qu'une pondération choisie a priori, y compris l'absence de pondération ce qui est habituellement utilisé. Si nécessaire, une seconde optimisation des premiers poids (associés aux premières variables latentes) peut être effectuée. L'ensemble des variables latentes multipliées par les pondérations associés peut être utilisé comme entrée d'un approximateur non linéaire tel que les RBFN (réseaux à fonctions radiales de base) et LSSVM.

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, indépendant, est uniquement utilisé pour évaluer les performances des algorithmes étudiés : on mesure celles-ci grâce à la NMSE (Normalized Mean Square Error). L'approximateur utilisé (LSSVM, [SUY 02]) comporte deux méta-paramètres (largeur des noyaux gaussiens et terme de régularisation) dont les valeurs optimales doivent être déterminées par comparaison de performances ; ceci est réalisé par une procédure de Leave-One-Out [SUY 02] sur l'ensemble d'apprentissage.

Quatorze variables latentes de la PLS sont sélectionnées par la procédure Leave-One-Out. Les variables latentes sont ensuite normalisées (moyennes nulles et variances unitaires) et la procédure de sélection des variables par Delta Test est utilisée ; elle résulte en la sélection des quatre premières variables latentes seulement.

La procédure de recherche des pondérations est également utilisée ; les résultats sont résumé dans la Table 1, pour une discrétisation k=10 ; 7 variables latentes sont retenues, au lieu de 4 pour la procédure de sélection. Les NMSE obtenus sur l'ensemble de test sont donnés dans la Table 2.

1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.8	0.6	0.3	0.3	0.4	0	0.2	0	0	0	0	0	0

Tableau 1: Poids résultant du Delta Test pour les 14 premières variable latentes.

Méthode de prédiction	NSME (test)
PLS (modèle linéaire avec 14 variables latentes)	0.0464
PLS + LSSVM (avec sélection des variables latentes par Delta Test)	0.0216
PLS + LSSVM (avec pondération des variables latentes de la PLS)	0.0045

Tableau 2: résultats des méthodes de prédiction sur la base de données Tecator.

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 et éviter le sur-apprentissage, la sélection et/ou la pondération des variables doivent être effectuée. La sélection ou la pondération des variables spectrales brutes est impossible dans des temps de calcul raisonnables. Par contre, la combinaison de cette méthode avec la construction de variables latentes par PLS est applicable dans des temps de calculs très courts (de l'ordre de 2 minutes de calcul sur un ordinateur personnel classique) .Les performances de la méthode sont illustrées sur la base de données Tecator ; elles sont supérieures à celles publiées précédemment dans la littérature [ROS 05].

Références

[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.

[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.

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

[JON 04] A. J. Jones, New Tools in Non-linear Modeling and Prediction. Computational Management Science, Vol. 1, Issue 2, p.p. 109-149, 2004.

[SUY 02] Suykens J.A.K., Van Gestel T., De Brabanter J., De Moor B., Vandewalle J., Least Vector Machines, World Scientific, Singapore, 2002 (ISBN 981- 238-151-1).

[ROS 05] F. Rossi, A. Lendasse, D. François, V. Wertz and M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modeling, Chemometrics and Intelligent Laboratory Systems, In Press.