Sélection supervisée d'attributs en haute dimension via l'approximation de Kirkwood

Claude Cariou, Kacem Chehdi

Institut d'Électronique et de Télécommunications de Rennes
Équipe TSI2M
Enssat
6, rue de Kerampont, 22300 Lannion, France
claude.cariou@univ-rennes1.fr, kacem.chehdi@univ-rennes1.fr

Résumé – Nous nous intéressons au problème de la sélection supervisée d'attributs, l'objectif étant de sélectionner séquentiellement et par ordre de pertinence les attributs permettant d'expliquer l'appartenance d'une observation à une classe particulière. Nous proposons une nouvelle technique de sélection séquentielle supervisée de type *filtre*, basée sur la maximisation de l'information mutuelle entre la vérité de terrain et les données multidimensionnelles. Cette technique s'appuie sur l'approximation de superposition de Kirkwood, qui permet d'appréhender le conditionnement multivarié nécessaire pour prendre en compte l'historique du processus de sélection. Nous illustrons la pertinence de notre approche par comparaison avec quelques méthodes similaires de sélection supervisée, notamment en imagerie hyperspectrale.

Abstract – We investigate the problem of supervised feature selection, with the objective to sequentially select the features which are the most informative and which best explain the membership of observations to particular classes. We propose a new filter-based forward feature selection method based on the maximization of the mutual information between the available ground truth and the multidimensional observations. Our technique uses the Kirkwood superposition approximation which allows to tackle the problem of multiple conditioning required to account for the history of the selection procedure. We illustrate the relevance of our approach by means of comparisons with similar feature selection methods, particularly in hyperspectral imagery.

1 Introduction

Nous nous intéressons au problème de la sélection supervisée d'attributs, utilisée dans de nombreuses applications, comme la reconnaissance de formes, la fouille de données, ou le domaine médical. La sélection d'attributs consiste à retenir, parmi un grand nombre d'attributs d'un ensemble de données, ceux dont l'apport informationnel est le plus grand. Il est fréquent en effet que plusieurs attributs d'un ensemble de données montrent une forte corrélation mutuelle, et que l'ensemble de tous les attributs disponibles n'apporte que peu d'information supplémentaire sur la structure des observations par rapport à un sous-ensemble d'attributs bien choisis. Selon le nombre d'observations disponibles, cette richesse informationnelle peut même nuire à la qualité d'un classification postérieure : c'est le phénomène de Hughes [1]. La sélection d'attributs peut être conduite de manière supervisée ou non supervisée. Ici, nous examinons les méthodes supervisées, et nous supposons qu'un ensemble labellisé d'observations multivariées est disponible. L'objectif est donc de sélectionner par ordre de pertinence les attributs permettant d'expliquer l'appartenance de telle observation à telle classe. On compte deux grandes familles de techniques de sélection supervisée d'attributs : l'approche par filtre (*filter*) dans laquelle la sélection est menée préalablement et donc indépendamment de la classification, et l'approche "symbiose" (*wrapper*) qui permet une sélection itérative par une évaluation en aval de la qualité de la classification obtenue au moyen des attributs courants [2, 3].

Dans cette communication, nous proposons une nouvelle technique de sélection séquentielle supervisée de type *filtre*, basée sur la théorie de l'information, qui propose un cadre approprié à la découverte de dépendances non linéaires, voire non fonctionnelles, entre variables. Notre approche est comparable à d'autres méthodes récentes [4, 5, 6, 7, 8] ayant pour objectif de maximiser l'information mutuelle entre la vérité de terrain et les données sélectionnées dans un schéma de type incrémental, c'est à dire par construction effective d'un sous-ensemble d'attributs pertinents. Cette technique s'appuie sur l'approximation de superposition de Kirkwood permettant d'appréhender le conditionnement multivarié nécessaire pour prendre en compte l'historique du processus de sélection. Après une description de notre approche, nous présenterons quelques résultats expérimentaux obtenus dans le contexte de la classification supervisée d'ensembles de données en grande dimension, notamment issues de l'imagerie hyperspectrale.

2 L'approche proposée : sAMMI

On considère une collection $\{\boldsymbol{x}_i = [x_i^{(1)} \cdots x_i^{(B)}]\}; i = 1, \cdots, N$ de réalisations d'un vecteur aléatoire réel $\boldsymbol{X} = \left[X^{(1)} \cdots X^{(B)}\right]$ de loi de densité supposée continue sur $\mathbb{R}^B$, avec B le nombre d'attributs originaux. On suppose connu un vecteur $\boldsymbol{y} = [y_1 \cdots y_N]^T$, $y_i \in C \subset \mathbb{N}$ de labels discrets associé à l'ensemble des observations. La sélection d'attributs consiste à trouver le vecteur $\boldsymbol{X}^\star$ de dimension $m < B$ qui permet d'expliquer au mieux la variable aléatoire des labels Y. Le problème peut être posé comme la maximisation de l'information mutuelle de Shannon entre la vérité de terrain et l'observation :

$$\boldsymbol{X}^\star = \arg \max_{\mathcal{P}_m(\{X^{(k)}\})} I(Y; \boldsymbol{X}) \quad , \tag{1}$$

avec :

$$I(Y; \boldsymbol{X}) = \sum_{y \in C} \int_{\mathbb{R}^B} f_{Y;\boldsymbol{X}}(y, \boldsymbol{x}) \log \frac{f_{Y;\boldsymbol{X}}(y, \boldsymbol{x})}{P_Y(y).f_{\boldsymbol{X}}(\boldsymbol{x})} \, d\boldsymbol{x} \tag{2}$$

et $\mathcal{P}_m(.))$ l'ensemble des parties de cardinal m d'un ensemble. Il existe de nombreuses méthodes supervisées de sélection séquentielle d'attributs dans la littérature [4, 5, 8]. Dans [4], les auteurs proposent un critère d'information mutuelle conditionelle (CMIM, pour *Conditional Mutual Information Maximization*) permettant de sélectionner l'attribut X' maximisant $I(Y; X'|X)$ pour chaque attribut X déjà sélectionné. Dans [5], la méthode mRMR (*min Redundancy Max Relevance*) fonde la sélection sur le critère $I(Y; X') - \frac{1}{\text{card}(\mathcal{S})} \sum_{X \in \mathcal{S}} I(X\,; X')$ où $\mathcal{S}$ est l'ensemble des attributs déjà sélectionnés. Dans [8], les auteurs proposent une autre méthode de sélection, appelée ici MIM (pour *Mutual Information Maximization*), maximisant un critère basé sur la réécriture du critère (1) en une somme de termes d'information mutuelle couplés et de termes d'information mutuelle conditionnelle. Ces trois méthodes partagent une hypothèse d'indépendance conditionnelle à un ordre supérieur à deux des attributs de l'ensemble de données. Le schéma que nous proposons, baptisé sAMMI (pour *superposition approximation Maximisation of Mutual Information*), repose sur la règle suivante de chaînage de l'information mutuelle multiple :

$$I(Y; X_1, \cdots, X_m) = \sum_{n=1}^{m} I(Y; X_n | X_1, \cdots, X_{n-1}) \; . \tag{3}$$

La difficulté de cette approche réside dans l'expression des termes d'information mutuelle multiplement conditionnelle, ceux-ci nécessitant théoriquement la manipulation de lois multivariées de dimension croissante.
Pour résoudre ce problème, nous proposons d'approcher des lois multiplement conditionnelles par l'approximation de superposition de Kirkwood (KSA, [9]). La KSA est utilisée en physique pour évaluer les distributions de conformations moléculaires en décomposant des distributions multivariées en grande dimension en distributions de dimension inférieure [10]. La KSA a également été utilisée en apprentissage automatique pour l'analyse des interactions entre variables [11]. Par exemple, l'approximation de superposition d'ordre deux (KSA2) d'une distribution trivariée est donnée par :

$$\begin{aligned} f_{X,Y,Z}(x,y,z) &\approx f^{\texttt{SA2}}_{X,Y,Z}(x,y,z) \\ &= \frac{f_{X,Y}(x,y) \; . \; f_{Y,Z}(y,z) \; . \; f_{X,Z}(x,z)}{f_X(x) \; . \; f_Y(y) \; . \; f_Z(z)}. \end{aligned} \tag{4}$$

Chaque terme d'information mutuelle conditionnelle $I^{\texttt{SA2}}(Y; X|X_1, \cdots, X_{n-1})$, $n > 1$ est estimé par échantillonnage suivant les distributions conditionnelles d'ordre inférieur, calculées par la KSA2. Plus précisément, étant donné un ensemble d'attributs antérieurement sélectionnés, sAMMI estime cette information mutuelle multiplement conditionnelle comme la moyenne empirique de termes d'information obtenus par échantillonnage de lois conditionnelles approchées par la KSA2 :

$$\begin{aligned} &I^{\texttt{SA2}}(Y; X|X_1, \cdots, X_{n-1}) \approx \\ &\tfrac{1}{M} \textstyle\sum_{i=1}^{M} \left\{ \sum_{y \in C} \int_{x \in \mathbb{R}} f^{\texttt{SA2}}_{Y;X|E^{(i)}_{n-1}}(y,x) \right. \\ &\left. . \log \frac{f^{\texttt{SA2}}_{Y;X|E^{(i)}_{n-1}}(y,x)}{P^{\texttt{SA2}}_{Y|E^{(i)}_{n-1}}(y).f^{\texttt{SA2}}_{X|E^{(i)}_{n-1}}(x)} \, dx \right\} \quad , \end{aligned} \tag{5}$$

où $f^{\texttt{SA2}}_{Y;X|E^{(i)}_{n-1}}$ est la KSA2 de la loi conjointe conditionnelle $f_{Y;X|X_1,\cdots,X_{n-1}}$ obtenue par échantillonnage, $E^{(i)}_{n-1}$ est l'évènement $\left\{X_1 = x_1^{(i)}, \cdots, X_{n-1} = x_{n-1}^{(i)}\right\}$, $n > 1$, $x^{(i)}$ représente le $i$ème échantillon issu de la loi univariée $f^{\texttt{SA2}}_{X|E^{(i)}_{n-1}}$, $P^{\texttt{SA2}}_{Y|E^{(i)}_{n-1}}$ et $f^{\texttt{SA2}}_{X|E^{(i)}_{n-1}}$ sont les lois marginales de $f^{\texttt{SA2}}_{Y;X|E^{(i)}_{n-1}}$, et M est le nombre de tirages issus de $f^{\texttt{SA2}}_{Y;X|E^{(i)}_{n-1}}$. L'Algorithme 1 montre la procédure d'estimation de cette information mutuelle conditionnelle, et l'Algorithme 2 décrit la sélection séquentielle d'attributs de sAMMI. En pratique, les distributions des marginales de $\boldsymbol{X}$ sont obtenues par discrétisation en 32 classes d'égale amplitude entre les valeurs extrêmes.

Le principal inconvénient de sAMMI réside dans sa complexité algorithmique qui croît avec le nombre de variables de conditionnement. Afin d'accélérer la sélection tout en préservant la capacité de sAMMI à prendre en compte des dépendances multiples dans les premières étapes de la sélection, nous en proposons également une extension permettant de basculer automatiquement vers la méthode CMIM, approche de sélection plus simple et plus rapide. Ce basculement est opéré dès que la décroissance relative (par itération) du terme d'information mutuelle conditionnelle moyenne atteint un seuil spécifié (5% ici). Cette extension est appelée sAMMI/CMIM dans la suite.

3 Résultats expérimentaux

Les méthodes sAMMI et sAMMI/CMIM ont été comparées avec les approches similaires citées plus haut [4,

Algorithme 1 SA2CONDMUTINFO

Initialisation :

N réalisations d'un ensemble de $n-1$ attributs sélectionnés $\mathcal{S} = \{X_1, \cdots, X_{n-1}\}$;

N réalisations d'un attribut candidat X ;

$\boldsymbol{y} = [y_1 \cdots y_N]^T$, la vérité de terrain correspondante, considérée comme un ensemble de réalisations de la v.a. Y ;

M, la dimension de l'échantillon aléatoire ;

Sortie : $I = I^{\texttt{SA2}}(Y; X|\mathcal{S})$

Estimer les lois 1-D $f_{X_k}, 1 \leq k \leq n-1$;

Estimer les lois 2-D f_{X_l,X_k}, $2 \leq k \leq n-1$, $1 \leq l \leq k-1$, et $f_{Y,X}(.,.)$;

Estimer les lois 3-D $f_{Y,X,X_k}, 1 \leq k \leq n-1$;

$I = 0$;

pour $i = 1 : M$ **{**

$x_1^{(i)} \sim f_{X_1}(x)$;

pour $k = 2 : n-1$ **{**

$$f^{\texttt{SA2}}_{X_k|E^{(i)}_{k-1}}(x) = \frac{\prod_{1\leq l\leq k-1} f_{X_l=x_l^{(i)},X_k}(x)}{[f_{X_k}(x)]^{(k-2)}} \frac{1}{K}$$

$$x_k^{(i)} \sim f^{\texttt{SA2}}_{X_k|E^{(i)}_{k-1}}(x)$$

}

$$f^{\texttt{SA2}}_{Y;X|E^{(i)}_{n-1}}(y,x) = \frac{\prod_{1\leq l\leq n-1} f_{Y,X,X_l=x_l^{(i)}}(y,x)}{[f_{Y,X}(y,x)]^{(n-2)}} \frac{1}{K'}$$

$$I \leftarrow I + .. \frac{1}{M} \sum_y \left\{ \int_x f^{\texttt{SA2}}_{Y;X|E^{(i)}_{n-1}}(y,x) \log \frac{f^{\texttt{SA2}}_{Y;X|E^{(i)}_{n-1}}(y,x)}{P^{\texttt{SA2}}_{Y|E^{(i)}_{n-1}}(y) \cdot f^{\texttt{SA2}}_{X|E^{(i)}_{n-1}}(x)} dx \right\}$$

}

Algorithme 2 SAMMI

Initialisation :

$\left\{ \boldsymbol{x}_i = [x_i^{(1)} \cdots x_i^{(B)}]; i = 1, \cdots, N \right\}$, un ensemble de réalisations de $\boldsymbol{X} = \left[X^{(1)} \cdots X^{(B)}\right]$;

$\boldsymbol{y} = [y_1 \cdots y_N]^T$, la vérité de terrain correspondante, considérée comme un ensemble de réalisations de la v.a. Y ;

Sortie : $\mathcal{S}$, l'ensemble des m attributs sélectionnés.

$\mathcal{R} = \left\{X^{(1)}, \cdots, X^{(B)}\right\}$, l'ensemble des attributs restants ;

$\mathcal{S} = \varnothing$, l'ensemble des attributs sélectionnés ;

$X_1 = \arg\max_{X\in\mathcal{R}} I(Y; X)$;

$\mathcal{S} \leftarrow X_1$; $\mathcal{R} \leftarrow \mathcal{R} \setminus X_1$;

pour $n = 2 : m$ **{**

$X_n = \arg\max_{X\in\mathcal{R}} I_{\texttt{SA2}}(Y; X|\mathcal{S})$; (Algorithme 1)

$\mathcal{S} \leftarrow \mathcal{S} \cup X_n$; $\mathcal{R} \leftarrow \mathcal{R} \setminus X_n$;

}

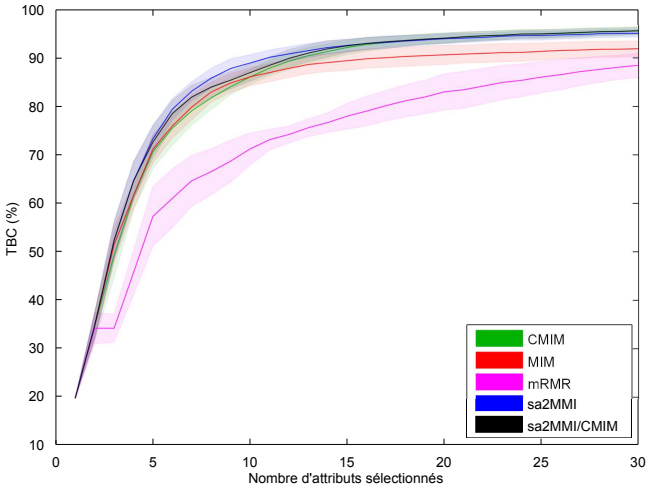

FIGURE 1 – Données MFEAT : Taux de bonne classification moyens et écart-types du classifieur 5-NN en fonction du nombre d'attributs sélectionnés pour différentes méthodes de sélection.

5, 8]. La méthodologie expérimentale consiste à comparer les taux de classification supervisée après sélection d'un même nombre de bandes spectrales par les différentes méthodes. Nous donnons ici les résultats pour deux ensembles de données. Pour chacun d'eux, 100 sous-ensembles d'apprentissage (50%) et de validation (50%) ont été extraits par échantillonnage stratifié afin d'obtenir moyennes et écarts-types des taux de bonne classification (TBC).

3.1 Données MFEAT

Le jeu de données MFEAT (*Multiple Features*) [12] comprend 2000 observations de 649 attributs correspondant à des descripteurs de chiffres manuscrits (10 classes, 200 observations par classe). Un prétraitement de binarisation des données a d'abord été effectué [5]. Les TBC moyens et écarts-types obtenus après sélection et classification sont donnés en Figure 1. On observe d'abord les faibles performances de MRMR au delà des deux premiers attributs sélectionnés. Un examen approfondi du coude des autres courbes montrent que SAMMI donne des TBC plus élevés que les méthodes CMIM et MIM (au niveau de confiance 95% après un test z unilatéral) entre 6 et 14 attributs sélectionnés. De plus, SAMMI reste meilleure que SAMMI/CMIM sur cette même plage, ce qui montre qu'elle peut, au prix d'une complexité accrue, proposer des premiers attributs plus pertinents que les autres approches.

3.2 Données AVIRIS

Nous avons ensuite appliqué notre approche au domaine de l'imagerie hyperspectrale pour la sélection de bandes spectrales les plus informatives. Dans ce contexte, la sélection de bandes spectrales est utile car elle permet une ré-

duction significative du volume de données tout en préservant l'information utile et nécessaire pour expliquer l'appartenance d'une observation à une classe donnée. Nous avons utilisé l'image AVIRIS *Indian Pines*, comprenant 220 bandes spectrales dans la gamme 0.4-2.5 μm [13]. Près de la moitié des pixels sont référencés dans une des 16 classes identifiées par une vérité de terrain. Après suppression des bandes les plus bruitées [14], nous avons retenu un ensemble de 185 bandes spectrales comme attributs initiaux pour chaque méthode. Pour chacun des 100 jeux d'apprentissage (50% apprentissage, 50% validation), les algorithmes CMIM, MIM, MRMR, SAMMI et SAMMI/CMIM ont produit un ensemble de bandes sélectionnées séquentiellement (de 1 à 30 bandes). Les données en dimension réduite obtenues pour chaque méthode ont ensuite fait l'objet d'une classification supervisée par 5-NN sur chaque jeu de validation. La Figure 2 montre l'évolution du taux de bonne classification global (moyenne $\pm$ écart-type) en fonction du nombre de bandes sélectionnées pour chacune des cinq méthodes de sélection. Les courbes montrent que les résultats obtenus avec SAMMI et SAMMI/CMIM sont les meilleurs en moyenne jusqu'à 10 bandes sélectionnées. Au delà de cette limite, les taux de classification sont dominés par la méthode CMIM, qui rejoint les taux donnés par la méthode SAMMI/CMIM. Pour cette dernière, le basculement à CMIM est effectif à partir de la huitième bande. Un test z unilatéral montre que les taux sont significativement plus élevés pour les méthodes SAMMI et SAMMI/CMIM que pour les trois autres méthodes (au niveau de confiance 95%) dans la gamme de 4 à 8 bandes sélectionnées, ce qui confirme l'apport de notre approche dans la sélection des tout premiers attributs. Des conclusions identiques ont été obtenues en remplaçant le classifieur 5-NN par un classifieur SVM à noyau gaussien, confirmant la supériorité de notre approche dans la sélection des premières bandes spectrales.

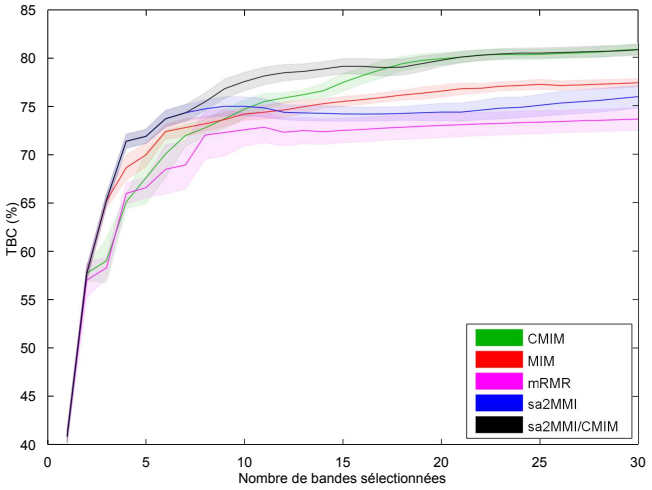

FIGURE 2 – Données AVIRIS : Taux de bonne classification moyens et écart-types du classifieur 5-NN en fonction du nombre d'attributs sélectionnés pour différentes méthodes de sélection.

4 Conclusion

Nous avons proposé une nouvelle technique de sélection supervisée incrémentale de type *filtre* basée sur la théorie de l'information, permettant d'appréhender le calcul d'une information mutuelle multiplement conditionnelle par l'approximation de superposition de Kirkwood. Cette approche, appelée SAMMI, met en œuvre l'estimation de lois conjointes du couple (attribut courant, classes d'apprentissage) par échantillonnage des lois conditionnelles aux attributs sélectionnés antérieurement. Elle a montré de meilleures performances dans la sélection d'attributs avant classification supervisée par rapport à des méthodes d'inspiration identique. D'un point de vue opérationnel, l'extension SAMMI/CMIM également proposée ici est intéressante, car elle permet d'atteindre de bons taux de classification tout en réduisant considérablement le coût calculatoire par rapport à SAMMI.

Références

[1] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[2] T. Lal, O. Chapelle, J. Weston, and A. Elisseeff, "Embedded methods," in *Feature Extraction : Foundations and Applications*, I. Guyon, S. Gunn, M. Nikravesh, and Z. L., Eds. Springer, Berlin, Germany, 2006, pp. 137–165.

[3] D. Erdogmus, U. Ozertem, and T. Lan, "Information theoretic feature selection and projection," in *Speech, Audio, Image and Biomedical Signal Processing Using Neural Networks*, B. Prasad and S. Prasanna, Eds. Springer, 2008, pp. 1–22.

[4] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, 2004.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information : criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[6] A. Martínez-Usó, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, 2007.

[7] S. Prasad and L. Bruce, "Decision fusion with confidence-based weight assignment for hyperspectral target recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1448–1456, 2008.

[8] B. Guo, R. Damper, S. Gunn, and J. Nelson, "A fast separability-based feature-selection method for high-dimensional remotely sensed image classification," *Pattern Recognition*, vol. 41, pp. 1653–1662, 2008.

[9] J. Kirkwood and E. Boggs, "The radial distribution function in liquids," *J. Chem. Phys.*, vol. 10, pp. 394–402, 1942.

[10] S. Somani, B. Killian, and M. Gilson, "Sampling conformations in high dimensions using low-dimensional distribution functions," *J. Chem. Phys.*, vol. 130, no. 13, pp. 134 102–, 2009.

[11] A. Jakulin, "Machine learning based on attribute interactions," Ph.D. dissertation, University of Ljubljana, Faculty of Computer and Information Science, 2005.

[12] [Online]. Available : http://archive.ics.uci.edu/ml/datasets/Multiple+Features

[13] [Online]. Available : http://cobweb.ecn.purdue.edu/~biehl/MultiSpec/documentation.html

[14] B. Mojaradi, H. Abrishami-Moghaddam, M. Zoej, and R. Duin, "Dimensionality reduction of hyperspectral data via spectral feature extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2091–2105, 2009.