Test de structure pour la prédiction de variable numérique

Stéphane Lallich – Fabrice Muhlenbach – Djamel A. Zighed

Laboratoire ERIC – Université Lumière – Lyon II
5 avenue Pierre Mendès-France, 69676 BRON Cedex – France
{lallich, fmuhlenb, zighed}@univ-lyon2.fr

RÉSUMÉ. *À l'occasion d'un précédent travail [ZIG 01], pour le cas de l'apprentissage par classement, lorsque la classe est catégorielle, nous avons développé un test de séparabilité faisant référence aux graphes de voisinage et aux outils d'analyse de l'autocorrélation spatiale. Dans ce papier, nous poursuivons cette démarche pour traiter le cas de l'apprentissage par la régression, lorsque la variable à prédire est numérique en utilisant le coefficient d'autocorrélation spatiale de Moran.*

MOTS-CLÉS : *Apprentissage supervisé, Classe continue, Graphe de voisinage, Autocorrélation spatiale, Coefficient de Moran*

1. Introduction

Nous abordons ici des problèmes d'apprentissage par la régression : la variable à prédire, ou *étiquette*, est numérique, les attributs sont de nature quelconque. Nous voulons construire un test de la capacité des attributs à fonder la prédiction de l'étiquette pour tout algorithme du type *distance-based*, ou test de structure. Comme le souligne Quinlan [QUI 92], l'apprentissage par la régression nécessite des méthodes spécifiques dans la mesure où les méthodes usuelles de l'apprentissage par classement (étiquette catégorielle) ne prennent pas en compte l'ordre des catégories créées après discrétisation de l'étiquette. Parmi ces méthodes (cf. [UYS 99]), pour le cas de l'apprentissage *distance-based*, nous citerons la méthode IBL [AHA 91] et la régression LWR [ATK 97].

Pour tester la structure d'une étiquette catégorielle, ou test de séparabilité, nous avons proposé une statistique de test associée au graphe de voisinage induit par les attributs : le *poids des arêtes coupées* (arêtes reliant des sommets de classe différente). Cette statistique fonde tout à la fois un test global pour tester la séparabilité de l'étiquette en fonction de l'information apportée par les prédicteurs [ZIG 01] et des tests locaux pour détecter les *outliers* [MUH 01]. Nous nous proposons de faire le même travail dans le cas de l'apprentissage par la régression. Prenons l'exemple de la base *Plasma* (section 5), où les attributs sont des facteurs de risque potentiel (régime, poids, etc.), l'étiquette étant la concentration plasmique en β-carotène (ou en rétinol pour la seconde étiquette). Nous allons tester le fait que la répartition des valeurs de l'étiquette sur les sommets du graphe de voisinage n'a pas de structure, ce qui rendrait illusoire tout apprentissage *distance-based*. Après un bref rappel sur les graphes de voisinage, nous présenterons l'autocorrélation spatiale (*ACS*) et le coefficient de Moran, puis le résultat de nos expérimentations.

2. Graphe de voisinage

Pour rendre compte de la proximité des exemples dans l'espace de représentation, à partir d'un simple indice de dissimilarité, nous utilisons les graphes de voisinage [ZIG 01] qui conduisent à des graphes planaires connexes et s'apparentent à des graphes de contiguïté spatiale. Nous avons choisi ici le Graphe des Voisins Relatifs [TOU 80] qui est un graphe connexe dans lequel deux points sont reliés par une arête lorsqu'ils vérifient la propriété $d(\alpha,\beta) \leq Max\left(d\left(\alpha,\gamma\right), d\left(\beta,\gamma\right)\right) \forall\gamma, \gamma \neq \alpha, \beta$ où $d(u,v)$ est une dissimilarité entre les exemples u et v. Deux points α et β sont reliés par une arête lorsqu'il n'y a pas d'autre point dans la *lunule* formée par ces deux points. Les n sommets du graphe de voisinage issu des p prédicteurs $X_j, j = 1, 2, ..., p$ correspondent aux exemples, alors que les a arêtes relient les exemples qui sont voisins relatifs. On note $\mathbf{V}$ la matrice d'incidence du graphe, $\mathbf{W}$ une matrice de poids associée aux arêtes (distances, rangs, etc.), éventuellement symétrisée en la moyennant avec sa transposée.

Suivant Cliff et Ord [CLI 81], on note $\sum_2 w_{ij} = \sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} w_{ij}$, et on adopte les notations suivantes : $S_0 = \sum_2 w_{ij}$, $S_1 = \frac{1}{2} \sum_2 (w_{ij} + w_{ji})^2$ et $S_2 = \sum_{i=1}^{n} (w_{i+} + w_{+i})^2$

3. Structure et autocorrélation spatiale

L'analogie entre graphes de voisinage et graphes spatiaux suggère le recours à l'analyse spatiale pour définir et tester la structure de la distribution d'une étiquette. En analyse spatiale, on considère des sites de $\mathbb{R}^2$ décrits par une variable Y. Il y a *ACS* positive lorsque les sites voisins présentent des valeurs semblables de Y, négative lorsqu'ils présentent des valeurs contradictoires. L'absence d'*ACS* signifie que la valeur présentée par un site ne dépend pas des valeurs présentées par les sites voisins. Pour tester l'absence d'*ACS* d'une variable numérique (H_0), on utilise classiquement les coefficients de Moran, Geary ou Getis. On peut adopter une présentation unifiée de ces coefficients en ayant recours à la statistique *cross-product* [MAN 67, HUB 76, GET 91].

La forme générale de la statistique *cross-product* pour l'analyse spatiale est : $\Gamma = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} U_{ij}$ où $\mathbf{W}$ est une matrice qui évalue la proximité spatiale des sites i et j, définissant ainsi un graphe spatial valué dont les sites sont les sommets, alors que U_{ij} mesure la proximité entre i et j suivant une autre dimension. Suivant le choix de U et à une normalisation près qui dépend de U mais pas de $\mathbf{W}$, on retrouve les coefficients usuels d'*ACS*.

Coefficient	notation	U_{ij}	dénominateur
Moran	I	$(y_i - \overline{y})(y_j - \overline{y})$	$\frac{2a}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$
Geary	C	$(y_i - y_j)^2$	$\frac{4a}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$
Getis et Boots	G	$y_i y_j$	$\sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j$

Alors que le C de Geary est du type ratio de Von Neumann, généralisant le coefficient d'autocorrélation temporelle de Durbin Watson, le I de Moran se présente comme le quotient d'une covariance locale par la variance totale. Sa valeur absolue peut exceptionnellement dépasser 1 lorsque la structure de poids est très déséquilibrée et que les valeurs extrêmes de Y sont affectées des plus forts poids. Sous l'hypothèse d'indépendance de Y et du graphe de voisinage, la covariance locale est nulle, ce qui implique la nullité de I. Le signe de I renseigne sur la nature positive ou négative de l'*ACS*. On peut remarquer que I est asymptotiquement le complément à 1 de C. On notera que le coefficient G n'a de sens que si Y est une variable métrique positive.

4. Test de structure à partir du coefficient de Moran

Nous avons choisi d'utiliser le I de Moran, considéré assez unanimement comme le meilleur choix, parce qu'il présente de bonnes propriétés locales [ANS 95] utiles pour des développements en cours (détection d'outliers) et qu'il est plus puissant que le C de Geary [CLI 81]. Sa variance serait moins sensible à la distribution des observations, b_2 (coefficient d'aplatissement de Pearson) y étant affecté d'un coefficient d'ordre n^{-2} et non n^{-1}.

Coefficient	Connexion : V	Pondération : W
Moran	$I = \frac{n}{2a} \frac{\sum_2 v_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$	$I = \frac{n}{S_0} \frac{\sum_2 w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$

Quelle que soit la forme retenue de la statistique *cross-product*, on a le choix entre deux schémas probabilistes. Le premier est le schéma gaussien *N* qui stipule que les observations sont le résultat de n tirages indépendants dans une population normale. Dans le second schéma *R*, moins restrictif, on raisonne conditionnellement aux n valeurs observées de Y et on suppose que la localisation de ces observations est le résultat d'un tirage au hasard pur parmi les $n!$ permutations de ces valeurs sur l'ensemble des localisations possibles.

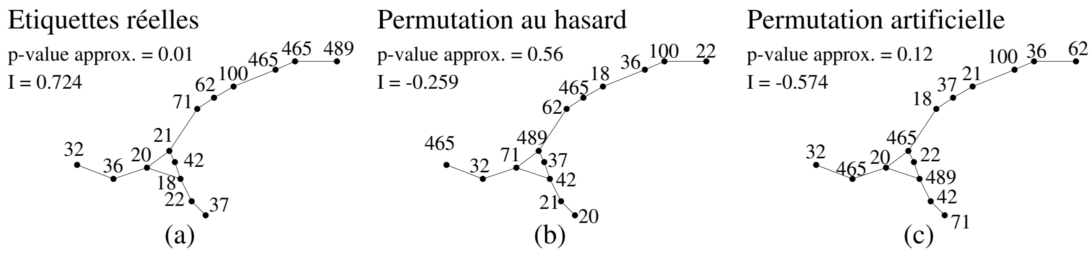

FIG. 1. *Illustration du graphe de voisinage et du test de structure*

On trouve dans [CLI 81] le calcul des moments relatifs à la loi suivie par I sous H_0, suivant que l'on se réfère au schéma *N* ou *R*. Sous des conditions peu contraignantes, la loi de I sous H_0 est asymptotiquement normale. La condition est que les quantités $\sum_{j=1}^{n}(w_{ij} + w_{ji})y_j$ ne soient pas trop différentes. Une condition suffisante pour assurer la normalité asymptotique est que $\max_i(\lambda_i^2/S_1)$ soit en $O(n^{-1})$, où λ_i désigne les valeurs propres de **W**. On peut ainsi calculer la *p-value* de I_{obs} par approximation normale. À défaut, sous le schéma R, on peut raisonner en statistique exacte si n est faible ou par simulation suivant la méthode de Monte-Carlo.

Coeff. c_m	schéma gaussien N	schéma randomisé R
$E(I)$	$\frac{-1}{n-1}$	$\frac{-1}{n-1}$
$E(I^2)$	$\frac{n^2S_1-nS_2+3S_0^2}{So^2(n^2-1)}$	$\frac{n\left[\left(n^2-3n+3\right)S_1-nS_2+3S_0^2\right]-b_2\left[\left(n^2-n\right)S_1-2nS_2+6S_0^2\right]}{(n-1)^{(3)}S_0^2}$

L'originalité de notre travail réside en l'association des graphes de voisinage et des coefficients d'*ACS* pour tester la structure d'une étiquette numérique indépendamment de l'algorithme d'apprentissage. Les travaux les plus proches des nôtres sont ceux liés aux applications des statistiques *cross-product* hors du champ spatial [HUB 76], notamment le gamma de Hubert en classification. On peut citer encore différents travaux, en analyse spatiale multivariée [LEB 69, THI 95, LEB 00] ou en économétrie des variables spatiales [ANS 93].

5. Expérimentations

Nous présentons d'abord un cas simple, le traitement d'un échantillon de 14 exemples extraits de *CPU*, une base du site de l'UCI [BLA 98]. Sur la figure 1, nous indiquons dans différents cas le graphe de voisinage et les étiquettes correspondantes, la valeur de I et sa *p-value* (R). La figure 1 (a) concerne les vraies étiquettes. Dans ce cas où $I = 0.724$ avec une *p-value* de l'ordre de 0.01, la distribution des étiquettes peut être considérée comme structurée, ce qui est illustré par l'ordonnancement des étiquettes sur le graphe. Sur la figure 1 (b), nous avons opéré une permutation aléatoire des étiquettes. I a alors une valeur faiblement négative mais non significative. En figure 1 (c), nous avons attribué les étiquettes à partir d'une permutation alternant les grandes et faibles valeurs de Y. Dans ce cas, $I = -0.574$ et la *p-value* = 0.12. La *p-value* n'est pas très éloignée de la signification mais la valeur négative de I indique qu'il s'agit d'un type particulier de structuration.

Nous avons ensuite appliqué (cf. Tab. 1) notre méthode à différentes bases du site de l'UCI (*auto-MPG*, *autos*, *CPU*, *housing* et *servo*) [BLA 98], ainsi qu'à une base du site StatLib de l'Université de Carnegie Mellon (*plasma*, avec deux variables à apprendre : β-*carotène* et *rétinol*) et à la base *pw-linear* générée suivant le modèle de Breiman [BRE 84]. Les diverses bases choisies pour notre étude ont toutes une étiquette numérique. Pour chaque base, nous avons retiré les individus présentant des données manquantes. Les éventuelles variables catégorielles ont été réencodées sous forme disjonctive complète. Les variables prédictives ont ensuite été centrées et réduites. Pour chaque base, en plus de la valeur du I de Moran et de sa *p-value*, nous indiquons le nombre d'individus présents dans la base (n), le nombre initial d'attributs prédictifs (p^*) et le nombre de variables prédictives numériques utilisées pour la constitution du graphe et le calcul des valeurs du test (p). Nous observons que pour l'essentiel des

Base	n	p^*	p	I	*p-value*	Base	n	p^*	p	I	*p-value*
auto-MPG	392	7	9	0.853	0	plasma (β-car.)	315	12	16	0.206	4.5E-07
autos	159	25	64	0.574	3.8E-18	plasma (rét.)	315	12	16	0.027	0.48
CPU (PRP)	209	6	6	0.581	0	pw-linear	200	10	10	0.327	5.4E-12
housing	506	13	13	0.743	0	servo	167	4	12	0.62	0

TAB. 1. *Tests de structure sur 8 bases*

bases d'apprentissage, le test de structure est très hautement significatif (*p-values* inférieures à 0.01). On pourra remarquer que la seule base pour laquelle ceci n'est pas le cas est *plasma* avec l'apprentissage de la variable *rétinol* alors que la distribution de β*-carotène*, qui comprend les mêmes individus et les mêmes valeurs pour les variables prédictives, est structurée. Pour ces deux variables, qui ne sont pas corrélées, les étiquettes de β*-carotène* se distribuent suivant un graphe du type présenté en figure 1 (a) alors que les étiquettes de la variables *rétinol* ressortent plutôt du type présenté en figure 1 (b).

6. Conclusion et travaux futurs

Cet article prolonge nos travaux sur les tests de séparabilité [ZIG 01] et nous permet maintenant de traiter aussi bien le cas d'une étiquette numérique que celui d'une étiquette catégorielle. À l'image de l'utilisation du test du poids des arêtes coupées pour traiter le bruit sur une classe catégorielle [MUH 01], nous avons entrepris d'utiliser le coefficient local de Moran pour traiter le bruit sur une classe numérique.

7. Bibliographie

[AHA 91] AHA D., KIBLER D., ALBERT M., Instance-based learning algorithms, *Machine Learning*, vol. 6, 1991, p. 37–66.

[ANS 93] ANSELIN L., Discrete Space Autoregressive Models, *GIS and Environmental Modeling*, 1993, p. 454–469.

[ANS 95] ANSELIN L., Local indicators of spatial association, LISA, *Geographical Analysis*, vol. 27, 1995, p. 93–115.

[ATK 97] ATKENSON G., MOORE A., SCHAAL S., Locally weighted learning, *AI Review*, vol. 11, n^o 1/5, 1997, p. 11–73.

[BLA 98] BLAKE C. L., MERZ C. J., UCI Repository of machine learning databases, Irvine, CA, 1998.

[BRE 84] BREIMAN L., FRIEDMAN J. H., OLSHEN R. A., STONE C. J., *Classification and Regression Trees*, 1984.

[CLI 81] CLIFF A. D., ORD J. K., *Spatial Processes : Models and Applications*, Pion Limited, London, 1981.

[GET 91] GETIS A., Spatial interaction and spatial autocorrelation : a cross products approach, *Env. & Planning A*, vol. 23, 1991, p. 1269–1277.

[HUB 76] HUBERT L., SCHULTZ L., Quadratic assignment as a general data analysis strategy, *British Journal of Mathematical & Statistical Psychology*, vol. 29, 1976, p. 190–241.

[LEB 69] LEBART L., Analyse statistique de la contiguïté, *Pub. de l'Inst. de Stat. de l'Univ. de Paris*, vol. 28, 1969, p. 81–112.

[LEB 00] LEBART L., Contiguïty analysis and classification, *Data Analysis*, Berlin, 2000, Springer Verlag, p. 233–244.

[MAN 67] MANTEL N., The detection of disease clustering and a generalized regression approach, *Cancer Res.*, vol. 27, 1967, p. 209–220.

[MUH 01] MUHLENBACH F., LALLICH S., ZIGHED D. A., Amélioration d'une classification par filtrage des exemples mal étiquetés, *ECA*, vol. 1, n^o 4, 2001, p. 155–166, Editions Hermes.

[QUI 92] QUINLAN J. R., Learning with Continuous Classes, *Proc. of AI'92*, Singapore, 1992, World Scientific, p. 343–348.

[THI 95] THIOULOUSE J., CHESSEL D., CHAMPELY S., Multivariate analysis of spatial patterns : a unified approach to local and global structures, *Environmental and Ecological Statistics*, vol. 2, 1995, p. 1–14.

[TOU 80] TOUSSAINT G. T., The Relative Neighborhood Graph of a Finite Planar Set, *Patt. Rec.*, vol. 12, 1980, p. 261–268.

[UYS 99] UYSAL I., GUVENIR H. A., An overview of regression techniques for knowledge discovery, *Knowledge Engineering Rev.*, vol. 14, 1999, p. 319-340.

[ZIG 01] ZIGHED D. A., LALLICH S., MUHLENBACH F., Séparabilité des classes dans R^{p}, *Actes SFC'01*, 2001, p. 356–363.