Deux méthodes de segmentation sur un Tableau de Données Symboliques[1]

J.P. Aboa Yapo[*] R. Emilion[†] F. Rossi[‡] B. Tang Ahanda[§]

Abstract

Nous présentons dans ce travail deux méthodes de segmentation sur un Tableau de Données Symboliques (TDS). Dans ce type de tableau, l'élément réel habituel du modèle tabulaire classique est remplacé par une variable aléatoire de carré intégrable. La nature du tableau induit une extension de la notion de question permettant le partitionnement. Nous proposons deux méthodes basées sur deux types de question, l'une aléatoire, et l'autre "symbolique". A partir d'une fonction d'impureté définie sur les noeuds de l'arbre binaire, on généralise ainsi le critère de Gini et la notion d'entropie, pour l'extension de l'algorithme CART [Breiman,84].

1 Problématique

On considère un ensemble C de N objets décrits par $p+1$ caractéristiques. Chaque objet c_i est associé à un $p+1$-uplet $(X_{i1}, ..., X_{ij}, ..., X_{ip}, Y_i)$ où X_{ij} est une variable aléatoire d'un espace probabilisé $(\Omega, \mathcal{F}, P)$ à valeurs réelles. Pour $j \in \{1, ..., p\}$, un descripteur X_j est une application de C dans $L^2(\Omega)$ et $X_{ij} = X_j(c_i)$ est la valeur prise par X_j pour l'objet c_i. Y_i est une réalisation d'une variable qualitative Y : c'est la variable à expliquer. Pour $i \in \{1, ..., N\}$ et $j \in \{1, ..., p\}$, Le tableau (X_{ij}) est appelé tableau de données symboliques, et introduit de manière naturelle les objets symboliques assertions présentés par Diday [Diday et al.,95].

[*]Lise-Ceremade, Université Paris IX-Dauphine, aboa@ceremade.dauphine.fr

[†]Centre de Calcul Informatique, Université de Dakar

[‡]Lise-Ceremade, Université Paris IX-Dauphine, rossi@ufrmd.dauphine.fr (Les coordonnées actuelles de Fabrice Rossi sont disponibles à l'URL `http://apiacoa.org/`.)

[§]Lise-Ceremade, Université Paris IX-Dauphine, ahanda@ceremade.dauphine.fr

[1]Publié dans les actes des Sixièmes journées de la Société Francophone de Classification. Disponible à http://apiacoa.org/publications/1998/sfc98.pdf

En analyse de données symboliques, le but de la segmentation est, à partir de l'ensemble C dont les objets sont représentés pour $i \in \{1, ..., N\}$ par $c_i = a_i \wedge b_i$, où a_i est un objet symbolique assertion dont la description est basée sur les variables aléatoires X_{ij} et b_i assertion de description basée sur la variable à expliquer Y, de rechercher un ensemble d'objets $e_t \wedge f_t$, représentant au mieux les objets de départ, avec e_t assertion décrivant le cheminement dans l'arbre correspondant à un noeud terminal t, et f_t description portée sur le résultat de l'affectation à une classe du nœud t.

Une des questions essentielles est sur un noeud donné, de pouvoir déterminer la question provoquant la coupure. En analyse de données classiques (X_j n'étant pas aléatoire), cette question en segmentation s'écrit sous la forme $[X_j > \alpha]$. Dans le cas d'un TDS, nous allons étendre cette question sous deux formes.

2 Segmentation sur une question aléatoire

Etant donné $\omega_k \in \Omega$ une observation, on obtient un tableau $(X_{ij}(\omega_k))$ de réalisations. L'idée est pour chaque observation, et pour une variable X_j donnée, de chercher des questions sous la forme :

$$\left[X_{ij}(\omega_k) > \alpha_j^i(\omega_k)\right] \tag{1}$$

α_j^i est donc une variable aléatoire, et à chaque observation, on peut lui associer une réduction aléatoire d'impureté :

$$\Delta I(\omega_k) = i_t(\omega_k) - \left\{P_{t_g}(\omega_k) i_{t_g}(\omega_k) + P_{t_d}(\omega_k) i_{t_d}(\omega_k)\right\} \tag{2}$$

où P_{t_g} (resp. P_{t_d}) est la proportion aléatoire du nombre d'éléments du noeud t qui tombent à gauche (resp. à droite) de la coupure. i_t est une variable aléatoire qui représente l'impureté, suivant qu'on considère le critère de Gini [Breiman,84] ou le critère lié à l'entropie [Quinlan,86]. On peut alors généraliser l'algorithme CART.

3 Segmentation sur une question symbolique

Dans cette section, un noeud est obtenu par la recherche de l'extension sur l'ensemble des objets symboliques initiaux d'un objet symbolique modal [Diday,97], à un seuil α par rapport à une fonction de comparaison $\mathcal{R} : L^2(\Omega) \times L^2(\Omega) \rightarrow [0\ \ 1]$. On cherche un descripteur X_j, un intervalle I, et une valeur α telle que la question posée à l'individu c_i soit :

$$[X_j\,(c_i)\,\mathcal{R}U_I] > \alpha \tag{3}$$

I est obtenu par subdivision de l'espace de description de la variable X_j en ensemble d'états, et U_I une variable aléatoire de loi uniforme sur I. Si $\mathcal{R}$ est une convolution on a :

$$\begin{aligned} X_j\,(c_i)\,\mathcal{R}1_I &= \int_\Omega X_{ij}\,(\omega) \circ U_I\,(\omega)\,dP\,(\omega) \\ &= \int_{I\!R} I\!I_I dP_{X_{ij}} \end{aligned}$$

Par suite

$$X_j\,(c_i)\,\mathcal{R}1_I > \alpha \Longleftrightarrow P\,(X_{ij} \in I) > \alpha \tag{4}$$

Le triplet (X_j, I, α) est trouvé de façon à maximiser la réduction d'impureté. On peut alors généraliser l'algorithme cité précédemment.

4 Conclusion

Le premier type de question ouvre des perspectives au sujet des propriétés de convergence des variables aléatoires intervenant dans la construction de l'arbre de décision.

La robustesse de la deuxième méthode devra être testée en utilisant des subdivisions de plus en plus fines, ou en utilisant d'autres fonctions de comparaison (Kolmogorov-Smirnov, etc.)

Nous implémentons ces méthodes avec un exemple de TDS issu de la biologie.

Bibilographie

[Baccini,75] Baccini A., Pousse A. (1975) "*Point de vue unitaire de la segmentation. Quelques conséquences*" - CRAS tome 280 série A p 241, Janvier 1975.

[Breiman,84] Breiman L., Friedman J.H., Olshen R. A., Stone C. J. (1984) "Classification and Regression trees " - Wadsworth & brooks 1984.

[Buntine,92] Buntine W., Niblett T. (1992) "*A further comparison of splitting rules for decision-tree-induction*" - Machine Learning, vol. 8, n° 1, January 1992

[Celeux et al.,82] Celeux G. et Lechevallier Y. (1990) "*Méthodes de segmentation. Analyse discriminante sur variables continues.*" ed. par Celeux G., p127 à 147.

[Cestnik et al. 87] Cestnik B., Kononenko I., Bratko I. (1987), "*Assistant 86: A knowledge-elicitation tool for sophisticated users*", Proceedings of the second European workshop on Machine Learning, 1987.

[Conruyt,94] Conruyt N. (1994) "*Amélioration de la robustesse des systèmes d'aide à la description, à la classification et à la détermination des objets biologiques.*" Thèse de doctorat, Université Paris IX-Dauphine. 1994

[Diday et al.95] Diday E. et Emilion R. (1995)"*Capacities and Credibilities in Analysis of Probabilistic Objets*" Proceedings of the International Conference on Ordinal and Symbolic Data Analysis - OSDA 95.

[Diday,97]Diday E. (1997)"*Extracting information from extensive datasets by symbolic data analysis*" Symbolic Data Analysis and its applications - CEFIPRA pp. 4-14, Université Paris IX-Dauphine, September 1997.

[Mingers,89] Mingers J. (1989) "*An empirical comparison of selection measures for decision tree induction*" - Machine learning, vol. 4, March 1989.

[Périnel,96] Périnel E. (1996) "*Segmentation et analyse de données symboliques. Le cas de données probabilistes*" Thèse de doctorat, Université Paris IX-Dauphine, 1996.

[Quinlan,86] Quinlan J. R. (1986) " *Induction of decision trees* " - Machine Learning, vol 1, n° 1, 1986.

[Quinlan,90] Quinlan J. R. (1990) "*Probabilistic decision trees*" Machine Learning éd. par Michalski R. S. et Kodratoff Y. pp. 140-152, San Mateo, CA: Morgan Kaufman, 1990

[Renyi,66] Renyi (1966) "Calcul des probabilités" - Dunod 1966

[Tangahanda,98] Tang Ahanda B. (1998) *"Extension de méthodes d'analyse factorielle sur des données symboliques"*, Thèse de doctorat, Université Paris IX-Dauphine.