Classification hiérarchique ascendante avec imputation multiple de données manquantes*

Ana Lorga da Silva
Laboratório de Estatística e Análise de Dados
Faculdade de Psicologia e Ciências da Educação, Universidade de Lisboa
Alameda da Universidade
1694-013 Lisboa Portugal
aigels@iseg.utl.pt

Gilbert Saporta
Chaire de Statistique Appliquée
Conservatoire National des Arts et Métiers
292 rue Saint Martin
75141 Paris cedex 03 France
saporta@cnam.fr

Helena Bacelar-Nicolau
Laboratório de Estatística e Análise de Dados (LEAD)
Faculdade de Psicologia e Ciências da Educação, Universidade de Lisboa
Alameda da Universidade
1694-013 Lisboa Portugal
hbacelar@fpce.ul.pt

RÉSUMÉ. En présence de données manquantes, et dans le cadre de la classification hiérarchique de variables on étudie deux méthodes utilisant les matrices obtenues après imputation multiple.
MOTS-CLÉS : Données Manquantes, Imputation Multiple, Classification Hiérarchique Ascendante

1. Introduction

Lorsque certaines valeurs sont manquantes dans un tableau de données, il est souvent nécessaire de les estimer avant d'appliquer une méthode statistique. L'imputation multiple [RUB 87] permet de restituer la variabilité des données par la substitution de m matrices de données complètes à une matrice comportant des données manquantes. Quand on fait ensuite

* Ce travail a été partiellement supporté par le Programme de Coopération Scientifique et Technique Luso-Française MSPLDM-542-B2 (Ambassade de France au Portugal et Ministère de la Science et de l'Enseignement Supérieur - ICCTI) co-dirigé par H. Bacelar-Nicolau e G. Saporta et par le Project d'Analyse des Données Multivariées (CEAUL-FCT) dirigé par H. Bacelar Nicolau.

une classification hiérarchique des variables, on cherche à obtenir à la fin une seule structure (consensus ou résumé) des m structures hiérarchiques obtenues.

2. Imputation Multiple, Matrices de similarité et Algorithmes de Classification

Soit m >1 matrices de données, $X^1, X^2, ..., X^m$, obtenues par application d'une méthode d'imputation multiple à une matrice avec données manquantes. Ces matrices ont ceci de commun que les n individus et les p variables sont les mêmes, la différence concernant un pourcentage i des observations, celles qui correspondent aux données manquantes de la matrice initiale.

Quand on utilise une méthode de classification sur les variables ou les individus, se pose alors la question: «Comment faire la combinaison des m matrices de façon à conclure sur la structure des données complètes?». On a développé deux méthodes qu'on comparera et évaluera en recourant à des matrices de données simulées.

2.1.1. *Première Méthode - Combinaison des matrices de similarité*

Cette méthode est basée sur la combinaison des matrices de similarité

Après avoir obtenu les m matrices par une méthode d'imputation multiple on détermine:

i. pour chaque matrice, $X^1, X^2, ..., X^m$, la matrice de similarité S_k, k=1,2,..., m,

ii. la moyenne des matrices de similarité S tel que, $S = \left(\sum_{k=1}^{m} S_k\right) \Big/ m$

iii. sur S, on utilise la méthode d'agrégation choisie,

iv. on détermine la structure hiérarchique correspondante, qu'on considérera représentative des m structures hiérarchiques correspondantes (associées à chaque $X^1, X^2, ..., X^m$).

2.1.2. *Deuxième Méthode - Combinaison des matrices des ultramétriques: «Consensus Ordinal IM»*

Cette méthode est basée sur la combinaison des structures hiérarchiques, considérant *l'ordre de l'agrégation*, (pas les niveaux d'agrégation), on la considère comme une méthode de consensus.

La procédure est la suivante:

i. pour chaque matrice, $X^1, X^2, ..., X^m$, on détermine sa matrice de similarité S_k, k=1,2,..., m,

ii. sur chaque matrice de similarité S_k, k=1,2,..., m, on utilise la méthode d'agrégation choisie, en obtenant pour chaque S_k une structure hiérarchique $H_k, k = 1, 2, ..., m$ $3 \le m \le 5$ (représenté par un dendrogramme),

iii. à chaque $H_k, k = 1, 2, ..., m$, est associée une matrice ultramétrique $U_k, k = 1, 2, ..., m$, qu'on détermine,

iv. on calcule les coefficients de Spearman r_s entre toutes les paires d'ultramétriques. On considère les cas où $r_s = 1$ qui correspondent à deux structures identiques. On cherche alors la structure ordinale majoritaire

v. le nombre de structures égales, n_i, doit être tel que

1) $n_i \in \left]\frac{m}{2}, m\right]$

2) si 1) n'est pas satisfaite on refait l'imputation on considérant m=10, avec le but de trouver un n_i satisfaisant 1) si ce n'est pas le cas, on dira qu'il n'y a pas d'arbre représentatif (pas de consensus).

C'est une condition semblable aux méthodes de consensus comme décrites par exemple en [GOR 99]. Si on obtient une hiérarchie représentative, on pourra parler d'une règle de consensus d'arbres de classification.

On appellera cette méthode «***consensus ordinal IM***».

2.2. *Application*

Comme dans des publications antérieures ([SIL 02] et [SIL 03]), on utilise des matrices de données complètes 1000×5, issues de 5 distributions multinormales correspondant aux structures suivantes

Figure 1 : Les 5 dendrogrammes

On enlève ensuite des données selon un modèle MAR - «Missing at Random».

On effectue ensuite 100 simulations de chaque cas.

On utilise comme coefficient de ressemblance le coefficient d'affinité, $c_a = \sum_{i=1}^{n} \sqrt{\frac{x_{ij}}{x_{.j}} \frac{x_{ij'}}{x_{.j'}}}$, où $x_{.j} = \sum_{i=1}^{n} x_{ij}$ et $x_{.j'} = \sum_{i=1}^{n} x_{ij'}$, défini par exemple en [BAC 85] et [BAC 02] et le coefficient de corrélation de Bravais-Pearson.

Comme méthodes d'agrégation on utilise les trois critères d'agrégation classiques: "average linkage", "single linkage" et "complete linkage".

On efface ensuite 10%, 15% et 20% des données de deux variables. Les données manquantes présentent un schéma majoritairement monotone - *«A monotone missing data pattern occurs when the variables can be ordered, from left to right, such that a variable to*

the left is at least as observed as all variables to the right» [STA 01] - avec un petit pourcentage de données manquantes représentées par un schéma non monotone.

On fait l'étude des résultats obtenus en utilisant les deux méthodes décrites, en utilisant la méthode d'imputation (m=5) pour toutes les données manquantes et en supprimant les lignes qui contiennent des données manquantes qui n'appartiennent pas au schéma monotone.

Le modèle d'imputation utilisé est basé sur la théorie Bayesienne [STA 01]. D'abord le modèle prédictif de régression OLS est estimé a partir des données complètes, comme d'habitude. On utilise ce modèle pour en générer d'autres où les valeurs des paramètres sont tirées au hasard dans la distribution *a posteriori*. "*The randomly drawn values are used to generate imputations, which include random deviations from the model's predictions*" ([STA 01]). De cette façon on garde plus de variabilité, car les paramètres sont estimés *a posteriori*.

Pour comparer les résultats aux structures originelles complètes on utilise le coefficient de Spearman entre ultramétriques.

Conclusion

Nous montrons sur nos simulations que les meilleurs résultats sont obtenus avec la première méthode associée au coefficient d'affinité avec les critères d'agrégation "average linkage" et "single linkage" et que le coefficient d'affinité est plus robuste que le coefficient de corrélation.

Bibliographie

[BAC 85] BACELAR-NICOLAU, H. "The Affinity Coefficient in Cluster Analysis", *Methods of Operation Research*, vol.53, 1985, p. 507-512, Martin J. Bekman *et al.* (ed), Verlag Anton Hain, Munchen.

[BAC 02] BACELAR-NICOLAU, H. "On the Generalised Affinity Coefficient for Complex Data", *Byocybernetics and Biomedical Engineering*, vol.22, n° 1, p. 31-42

[GOR 99] GORDON, A.D. *Classification*, Chapman & Hall, 1999.

[LIT 87] LITTLE, R. J. A. & RUBIN, D. B. *Statistical Analysis With Missing Data*, John Wiley & Sons, New York, 1987.

[RUB 87] RUBIN, D.B. *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, 1987.

[SIL 02] SILVA, A.L, BACELAR-NICOLAU, H. & SAPORTA, G. "Missing Data in Hierarchical Classification of Variables - a Simulation Study" in *Classification Clustering and Data Analysis*, 2002, p.121-128, Springer.

[SIL 03] SILVA, A.L, BACELAR-NICOLAU, H. & SAPORTA, G. "Efeito de um Método de Imputação Múltipla em Classificação Hierárquica de Variáveis", *Proceedings JOCLAD 2003 X Jornadas de Classificação e Análise de Dados*, 2003 p. 130-142

[STA 01] STATISTICAL SOLUTIONS, Lda. "*SOLAS for Missing Data Analysis*, 3.0". Cork, Ireland: Statistical Solutions, 2001