Missing Data in Hierarchical Classification – a study with Personality developpment data

Ana Lorga da Silva[1], Helena Bacelar-Nicolau[2], Gilbert Saporta[3], Manuel Geada[4]

[1]ISEG, Universidade Tecnica de Lisboa
e-mail:aigcls@iseg.utl.pt
[2]LEAD-FPCE, Universidade de Lisboa
e-mail:hbacelar@fpce.ul.pt
[3]Chaire de Statistique Appliquée
Conservatoire National des Arts et Métiers, Paris, France
e-mail:saporta@cnam.fr
[4]FPCE, Universidade de Lisboa
e-mail:geadamlc@fpce.ul.pt

In this work we analyse the effect of missing data in hierarchical classification of variables according to the following factors: amount of missing data, imputation techniques, similarity coefficient, and aggregation criterion. We have used two methods of imputation, a regression method using an ordinary-least squares method and an EM algorithm. For the similarity matrices we have used the (unweighted) basic affinity coefficient $c_a = \sum_{i=1}^{n} \sqrt{\frac{x_{ij}}{x_{.j}} \frac{x_{ij'}}{x_{.j'}}}$, where $x_{.j} = \sum_{i=1}^{n} x_{ij}$ and $x_{.j'} = \sum_{i=1}^{n} x_{ij'}$, as defined for instance in Bacelar-Nicolau(2000).and the Bravais-Pearson correlation coefficient

In this work we are interested in the classification of variables. We use the following hierarchical aggregation criteria as defined in Anderberg(1973):

Average linkage $C(A,B) = \frac{1}{(\#A)\times(\#B)} \sum c(X_j, X_{j'}),\ X_j \in A, X_{j'} \in B$.

Single linkage $C(A,B) = max\{c(X_j, X_{j'}), X_j \in A, X_{j'} \in B\}$ and

Complete linkage (CL): $C(A,B) = min\{c(X_j, X_{j'}), X_j \in A, X_{j'} \in B\}$ where A and B represent two clusters and c is a similarity coefficient between two variables (X_j , $X_{j'}$ are ($n \times 1$) variables).

In order to compare hierarchical classification models, we will use the Spearman's coefficient – c_s - between the ultrametic matrices, based on pairs of observations with the usual correction for ties.

In the present study we use a set of real data – Personality development data (Geada, M.(1998)) - under the reference hypothesis that the data are issued from a multinormal population (Saporta(1990)).

The missing data problem has been dealt in a large number of papers and books where several methods to minimise missing data effect have been developed (Little and Rubin(1987), Dempster, Laird and Rubin(1977), among others).

In this work we are particularly interested in the data missing at random. The expression of the general notion of MAR can be then written as: $Prob(R|X_{obs},X_{mis}) = Prob(R|X_{obs})$, where X_{obs} represents the observed values of $\mathbf{X}_{n\times p}$, X_{mis} the missing values of $\mathbf{X}_{n\times p}$ and $R = \lfloor R_{ij} \rfloor$ is a missing data indicator, $R_{ij} = \begin{cases} 1, & \text{if } x_{ij} \text{ observed} \\ 0, & \text{if } x_{ij} \text{ missing} \end{cases}$

We have been studing –whith simulation experiments - the performance of the affinity and the Pearson's correlation coefficients as measures of similarity between variables, in hierarchical classification in presence of missing data, and when the missing data are filled-in using the two imputation methods as mentioned. Here we are interested to do a similar study over the observed data. This study with the real data is not exactly the same as the simulation experiments, the second one deals with five variables (with multinormal distribution), the data is missing over two variables - 10%, 15% and 20% (over the total of the data – each of the 1000×5 matrices) – 25%, 37,5% and 50% of missing data (MD) over two variables. Here we have seven variables (a 181x7 matrix) and we take of the data on one variable (25%, 37,5% and 50% of MD).

When one analysis the structure of the complete data we find two main groups – "objective behaviour"/"cognitive emotional behaviour" - {transgressive, delinquency} and {nurturance, health separation, selfconcept, family function, Coping skills} as it was expected, in terms of psycological dimentions. The obtained results were also mathematicaly expected - the basic afinity coefficient revels once more rubuster than the correlation coefficient, the structure obtained using the basic affinity coefficient is exactly the same for the three methods (with the complete data), while using the Pearson's coefficient the hierarchical classification structures are not the same, but the ultrametrics are "significantly correlated", also better results are obtained in presence of MD using the basic affinity coefficient.

Keywords: Missing Data, Hierarchical Cluster Analysis, Affinity Coefficient, Pearson's Coefficient, Spearman's Coefficient, Ultrametric, OLS method, EM Algorithm, Personality development.

Main References

ANDERBERG, M. R. (1973) Cluster analysis for applications, Academic Press, New York.

BACELAR-NICOLAU(2000) The Affinity Coefficient in Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data. H.H. Bock and E.Diday (Eds.), Springer,160-165.

BEALE, E. M. L. and LITTLE, R. J. A.(1975) Missing values in multivariate data analysis. J. R. *Statist. Soc.* B, **37**, 129-145.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B.(1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc*. B, **39**, 1-38.

GEADA, M. (1998), Family functioning, personality development and deviant peer group membership on adolscence: Implications for drug use prevention, Proceedings Workshop EWODOR, Ed. EWODOR

LITTLE, R. J. A. and RUBIN, D. B.(1987) Statistical Analysis With Missing Data, John Wiley & Sons, New York.

SAPORTA, G.(1990) Probabilités, Analyse des Données et Statistique, Editions Technip, Paris