A non parametric pre-grafting procedure for data fusion

Massimo Aria
Department of Mathematics and Statistics
University of Naples Federico II
aria@unina.it

Antonio D'Ambrosio
Department of Mathematics and Statistics
University of Naples Federico II
antdambr@unina.it

Abstract: *Before proceeding with the fusion, which is a very special case of missing data imputation, pre-fusion conditions should be verify In this paper we propose a procedure to verify the presence of a common data structure between two data files based on non-parametric bootstrap confidence intervals.*

Keywords: Data Fusion, Bootstrap, Classification and Regression Trees, Pre-Fusion Conditions

Introduction

Data Fusion and Data Grafting are concerned with combining files and information coming from different sources [Saporta (2002)]. The problem is not to extract data from a single database, but to merge information collected from different sample surveys.
The term fusion is used in this sense. The typical data fusion situation formed of two data samples, the former made up of a complete data matrix **X** relative to a first survey, and the latter **Y** which contains a certain number of missing variables.
The aim is to complete the matrix **Y** beginning from the knowledge acquired from the **X**. As a consequence, the Data Fusion can be considered as a particular case of data imputation framework, with the difference that in this case a group of instances is missing as they have not been collected.

Data Fusion framework

Data Fusion problem can be formalized in terms of two data files [Aluja-Banet *et al*. (1998)]. The first data file consists of a whole set of $p + q$ variables measured on n_0 individuals. This data file is called *donor file*. The second data file, usually named *receptor file*, consists in a subset of p variables measured on n_1 units.
The problem is to merge two different and independent databases.

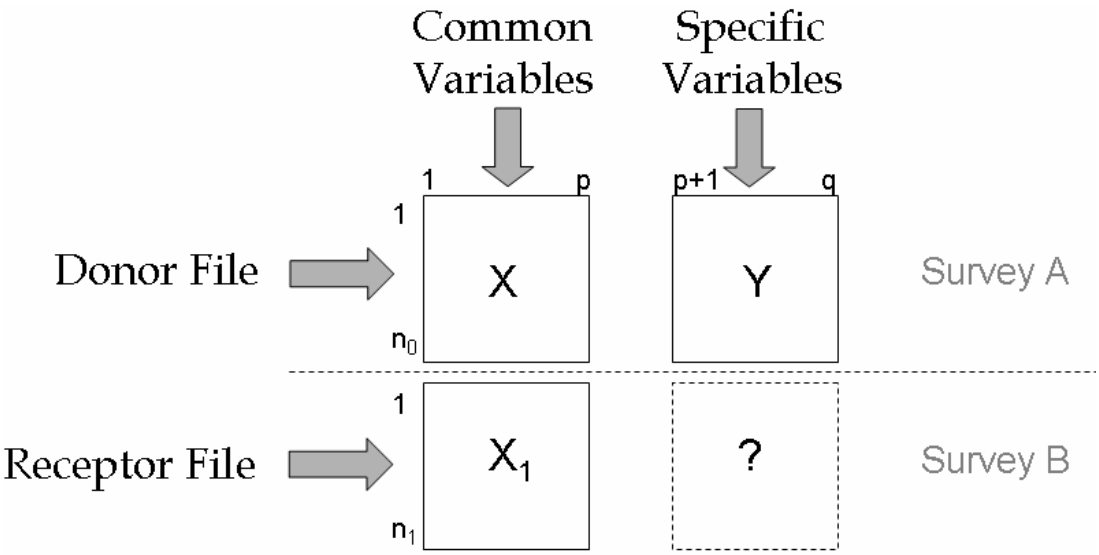

Figure 1: Data Fusion mechanism

In literature, usually on distinguish between explicit models and implicit models.
With **explicit models**, a model is used to connect **Y** variables with the **X** variables in the donor file and then applying this model in the receptor file.
Within explicit models on distinguish between the *classical approach* (regression models, general linear models and logistic regression) and the *non-parametric approach* (Tree-based methods) [D'Ambrosio *et al.* (2007)].
Implicit models are based on the concept of similarity among the observations deriving from different sources [Alujia-Banet *et al.*(1998, 2007)].

Pre-Fusion conditions

Before proceeding with the fusion, pre-fusion conditions should be verified, indeed the internal relationships of common variables **X** and $\mathbf{X_1}$ should show a stable pattern [Bonnefous *et al.*(1986)]. Rius *et al.* [Rius *et al.* (1999)] suggest to verify initial conditions by identifying the common group of variables which define a similar representation subspace for both data files.
Authors chose the common variable space from the different data sets by performing a Principal Component Analysis on **X** and then they use a branch-and-bound procedure to eliminate variables in order to find a minimal set of variables of the common group. In a succeeding step, they analyze the stability of the common space by bootstrap replications to assure that the association of the common variables is the same.
We propose a methodology to verify pre-fusion conditions which is entirely based on the use of tree-based models.
Let $x_1,\ldots,x_p$ be the set of common variables belonging to the donor file **X**, and let $x_1^1,\ldots,x_p^1$ be the set of common variables belonging to the receptor file $\mathbf{X_1}$.
We propose to build up p classification (or regression) trees considering, at turn as response variable, the j^{th} x variable validated by cross-validation. For each tree we compute a confidence

interval for the misclassification ratio (or for the root mean squared error) through *k* bootstrap replications. Then the common variables forming the receptor file are dropped down the *p* trees.
For each variable, if the goodness of fit measure is included in the confidence interval then the two data files (both donor and receptor) should have a stable pattern. In addition, this procedure is an alternative way to select the suitable number of variables which have to be used for the fusion process.

Simulation study

To test the methodology, several simulation studies were performed. Following tables show the results relative to only one simulated data set.
Simulation study has been defined thinking to reliable situations in which Data Fusion can be functional, i.e. when the donor file is a set of socio-economic variables (i.e., age, gender, income, job, etc.). For that reason, a simulated dataset was built using different random distributions for the set of common variables (Uniform, Multinomial, Normal), whereas specific variables were generated in both cases without relationship with common variables and with linear link with other variables (see table 1). Entire data set was randomly splitted in two sub-sets (donor and receptor file), used in the pre-grafting procedure.

Simulation Study	
Donor file: 1000 observations; Receptor file: 400 observations;	
Common Variables	*Specific Variables*
X_1 is uniform in $\{18,65\}$	$Y_1 = k + 0.4X_4 - 0.2X_1$
X_2 is multinomial $(\pi = \{0.2,0.4,0.1,0.3\})$	$Y_2 = k - 0.3X_5 + 0.1X_4 - 2X_2$
X_3 is multinomial $(\pi = \{0.2,0.5,0.3\})$	$Y_3 \sim N(X_1 - X_4) + \exp(0.7X_3 + 0.3X_1)$
X_4 is normal $(\mu = 100, \sigma = 10)$	Y_4 is uniform in $\{0,100\}$
X_5 is normal $(\mu = 500, \sigma = 50)$	

Table 1: Simulation study

Table 2 contains the bootstrap confidence intervals for the root mean squared error (variables x_1, x_4 and x_5) and the misclassification ratio (variables x_2 and x_3) of the variables belonging to the common part of donor file. Confidence interval has been derived using the percentile approach with $\alpha = 0.05$.
Table 3 shows the performance of the decision trees validated via cross-validation of the same variables belonging to the receptor file.

	x_1	x_2	x_3	x_4	x_5

lower limit	6,813	0,137	0,113	5,285	27,613
upper limit	12,846	0,300	0,297	9,968	49,282

Table 2: *Bootstrap Confidence Intervals*

x1	x2	x3	x4	x5
RMSE	MR	MR	RMSE	RMSE
12,295	0,280	0,240	9,071	46,122

Table 3: *Decision tree badness of fit measures*

These results prove that both donor and receptor file have a similar data structure, which is a fundamental pre-condition of the Fusion process.

References

Aluja-Banet T., Rius R., Nonell R., Martìnez-Abarca, M.J. (1998) Data Fusion and File Grafting. *Analyses Multidimensionelles Des Données*, NGUS 97. 1 ed. Parìs: CISIA-CERESTA, Eds. A. Morineau, K. Fernàndez Aguirre, P. 7-14.

Aluja-Banet T., Daunis-i-Estadella J., Pellicer D. (2007) GRAFT, a complete system for data fusion. *Computational statistics & Data analysis* 52, 635-649.

Barcena, M.J., Tusell, F. (1999). Enlace de encuestas: una propuesta metodològica y aplicaciòn a la Encuesta de Presupuestos de Tempo. *Questiio*, vol. 23, n\'{u}m. 2, pp. 297--320.

Bonnefous S., Brenot J., Pagés J.P. (1986). Mèthode de la greffe et communication entre enquetes. *Data analysis and Informatics IV*.

D'Ambrosio A. (2008). *Tree based methods for Data Editing and Preference Ranking*. PhD thesis, Department of Mathematics and Statistics, University of Naples Federico II, Napoli.

D'Ambrosio A., Aria M., Siciliano R. (2007). Robust Tree-based Incremental Imputation Method for Data Fusion. *Advances in Intelligent Data Analysis*, Springer-Verlag, pp 174-183.

Efron B., Tibshirani R.J. (1979). *An introduction to the Bootstrap*. Chapman and Hall.

Rius R., Nonell R., Aluja-Banet T. (1999). File grafting in market research. *Applied stochastic models in business and industry* 15, 451-460.

Saporta G. (2002) Data fusion and data grafting. *Computational Statistics & Data Analysis* 38, 465-473.

