Chapitre 4 : Problème de données manquantes dans un tableau numérique. Une application de l'A.S.I.

Régis Gras

Equipe COnnaissances & Décision (COD)
Laboratoire d'Informatique de Nantes Atlantique – FRE CNRS 2729
Site Ecole Polytechnique de l'Université de Nantes
La Chantrerie BP 50609 44306 Nantes cedex 3
regisgra@club-internet.fr

Résumé. Une base de données croisant variables et sujets issues d'observations présente souvent des vides dus, par exemple, à des absences de réponse ou à l'impossibilité matérielle de la recueillir. Or, pour en effectuer un traitement, il est essentiel de disposer d'un tableau complet. L'analyse statistique implicative, entre autres méthodes d'analyse de données, à l'œuvre au moyen du logiciel de traitement C.H.I.C. (Couturier et Gras., 2005), impose que les vides soient comblés. Se pose alors le problème de déterminer quelle valeur la plus vraisemblable attribuer à la variable non observée sur tel sujet ou, de façon symétrique, quelle valeur attribuer à un sujet sur une variable donnée et muette sur lui. Nous présentons ici une méthode qui, au vu du comportement de réponse observé par le sujet sur d'autres variables, intimement liées à la variable muette, permet de pallier la carence locale. Un exemple numérique illustre l'usage de cette méthode sur un tableau incomplet.

1 Problématique et contraintes sémantiques

L'analyse statistique implicative traite un tableau de données qui croise un ensemble de variables V, colonnes du tableau, de cardinal v et un ensemble E de sujets (ou d'objets), lignes du tableau, de cardinal n. L'intersection d'une ligne et d'une colonne représente donc la valeur x(i) prise par l'individu x selon la variable i. Dans le cas où les variables V sont binaires x(i) = 0 ou 1. Dans les autres cas classiques (variables modales, fréquentielles, numériques normées), les valeurs sont des nombres réels de l'intervalle [0,1].

Or il est possible que les observations ou les mesures faites conduisant au tableau présentent des "trous", c'est-à-dire des absences de réponse. Il existe donc des individus x et des variables i telles que la **valeur x(i) soit manquante**. Si l'on ne souhaite pas supprimer l'individu x, donc toutes les autres observations faites en x selon les autres variables, et par suite perdre un certain nombre d'informations permettant d'étudier les relations implicatives entre variables, le problème va consister à choisir la valeur la plus pertinente que l'on pourra affecter à x(i) pour la structure des variables sous-jacente.

L'idée intuitive est de chercher quels sont les individus qui ont des comportements semblables à x selon les autres variables sur lesquelles nous disposons d'informations pour x et d'attribuer à x(i) la valeur correspondant à celle que prend en i l'individu y, le plus "ressemblant" à x. La modélisation va donc porter sur le choix d'un critère de ressemblance entre individus.