Classification de variables en présence de valeurs manquantes : application aux données de préférence

Karin Sahmer, Evelyne Vigneau et El Mostafa Qannari

Laboratoire de sensométrie et de chimiométrie,
ENITIAA / INRA, rue de la Géraudière, BP 82 225,
44322 Nantes Cedex 03
e-mail sahmer@enitiaa-nantes.fr (Karin Sahmer)

Dans les industries agro-alimentaires, l'étude des préférences des consommateurs est primordiale pour le développement et l'amélioration des produits. Dans le cadre des tests de préférence, il est d'usage de demander à un panel de consommateurs de donner des notes de préférence à une gamme de produits. Très généralement, cette épreuve se traduit par l'existence de plusieurs groupes de consommateurs ; chacun des groupes exprimant une préférence pour un ensemble de produits et un rejet pour d'autres. L'analyse des données de préférences doit exhiber ces groupes et en tenir compte dans les différentes phases de l'étude. Dans cette perspective, on effectue une classification des consommateurs sur la base des notes attribuées aux différents produits. Pour cela, on peut utiliser une méthode de classification de variables, qui a été proposée par Vigneau et Qannari (2003).

Cependant cette méthode requiert que les données soient complètes, ce qui présente une sérieuse limitation dans la pratique. En effet, lorsque le nombre de produits est relativement important, il est d'usage de recourir à un plan de dégustation incomplet pour des raisons de coût, de temps ou de contraintes matérielles. Il est par conséquent utile d'adapter la démarche de classification à la situation où chacun des consommateurs n'évalue qu'une partie des produits. La méthode la plus simple consiste à remplacer chaque valeur manquante par la moyenne du consommateur ou du produit. La classification est ensuite effectuée comme si on avait affaire à des données complètes. Les résultats peuvent être améliorés par un renouvellement des imputations dans les groupes définis par la classification (Sahmer, 2003).
Les performances des différentes méthodes sont comparées à l'aide de simulations.

Mots-clés : algorithme de partitionnement, classification de variables, classification hiérarchique, données de préférences, données manquantes

Références

SAHMER, K. (2003) : *Classification des variables en présence de données manquantes : Application* aux *données de préférence.* Diplomarbeit, Fachbereich Statistik, Universitat Dortmund.
VIGNEAU, E. and QANNARI, E.M. (2003) : Clustering of variables around latent components. *Communications in Statistics — Simulation and Computation, 82, 1131-1150.*