Classification de variables avec possibilité de mettre à l'écart des variables atypiques ou de bruit. Implémentation dans le package ClustVarLV.

Evelyne Vigneau*[1] and El Mostafa Qannari

[1]Oniris, Unité de Sensométrie et Chimiométrie (Ecole Nationale Vétérinaire, Agro-alimentaire et de l'Alimentation Nantes Atlantique) – Ministère de l'alimentation de l'agriculture et de la pêche – France

Résumé

En général, les méthodes de classification ne permettent pas de tenir compte de la présence de données atypiques ou de bruit. Le processus de classification concerne tous les objets statistiques, même ceux qui n'ont pas de liens (proximités) avec aucun autre objet du jeu de données. Les méthodes de classication floue offrent la possibilité d'affecter ce type d'objets à différentes classes mais avec un degré d'appartenance faible. On adoptera des approches de classification "nette", mais en offrant la possibilité qu'un objet atypique ne soit affecté à aucune classe, ou, alternativement, que cet objet ne soit pas pris en compte dans la dénifition des noyaux des différentes classes. On se placera dans le cadre de la classification de variables, les objets statistiques considérés pouvant être des items de réponse d'un questionnaire; des caractéristiques sensorielles, biochimiques; ou encore des variables spectométriques ou chromatographiques quantiées par des approches holistiques de type -omiques, etc.
La méthode de classification de variables autour de variables latentes (CLV) [1], implémentée dans le package ClustVarLV [2], a été adaptée an de pouvoir mettre à l'écart les variables assimilables à du bruit ou atypiques en regard de la structure des corrélations au sein du jeu de données. A partir d'une partition initiale des variables, l'algorithme d'optimisation alternée de CLV consiste en deux étapes : (a) une étape d'estimation de la variable latente de chaque groupe, et (b) une étape de (ré-)affectation de chacune des variables au groupe avec lequel elle présente le lien le plus fort. Dans la méthode CLV, selon l'objectif poursuivi par l'utilisateur, ce lien est mesuré soit par la covariance, soit par la covariance au carré, entre la variable considérée et la composante latente du groupe. Deux stratégies différentes de modification de l'algorithme sont proposées.
...

Il est possible de mettre en oeuvre, l'une ou l'autre de ces deux stratégies en utilisant la fonction CLV_kmeans() du package ClustVarLV. Deux exemples de mise en oeuvre de cette démarche de classication de variables, avec "nettoyage", seront brièvement présentés. Dans le contexte de la segmentation de consommateurs en évaluation sensorielle, la possibilité de mettre de côté des consommateurs dont les préférences ne s'accordent pas avec les grandes tendances au sein du panel, permet d'exhiber une segmentation plus interprétable. Dans le domaine des techniques analytiques à haut débit, telle que la RMN, qui permettent de collecter un grand nombre de points de mesure pour chaque échantillon, la classification CLV permet d'identifier des sous-ensembles de variables fortement corrélées. Les stratégies

*Intervenant

complémentaires proposées ici permettent, également, d'écarter les variables spectrales dont la variabilité serait assimilable à du bruit.

Références

Vigneau, E., Qannari, E. M. (2003). Clustering of variables around Latent Variables. Comm.Stat. - Simul. Comput., 32(4), 1131-1150.

Vigneau, E., Chen, M. (2015). ClustVarLV: Clustering of variables around Latent Variables.URL http://CRAN.R-project.org/package=ClustVarLV,R package version 1.3.2.

Davé, R. N. (1991). Characterization and detection of noise in clustering. Pattern Recognition Letters, 12,(11), 657-664.

Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse principal component analysis. J. Comput. Graph. Statist., 15, 265-286.