Application du coclustering à l'analyse exploratoire d'une table de données

Aichetou Bouchareb*, Marc Boullé*, Fabrice Clérot*, Fabrice Rossi**

*Orange Labs
prenom.nom@orange.com
**SAMM EA 4534 - Université Paris 1 Panthéon-Sorbonne
prenom.nom@univ-paris1.fr

Résumé. La classification croisée est une technique d'analyse non supervisée qui permet d'extraire la structure sous-jacente existante entre les individus et les variables d'une table de données sous forme de blocs homogènes. Cette technique se limitant aux variables de même nature, soit numériques soit catégorielles, nous proposons de l'étendre en proposant une méthodologie en deux étapes. Lors de la première étape, toutes les variables sont binarisées selon un nombre de parties choisi par l'analyste, par discrétisation en fréquences égales dans le cas numérique ou en gardant les valeurs les plus fréquentes dans le cas catégoriel. La deuxième étape consiste à utiliser une méthode de coclustering entre individus et variables binaires, conduisant à des regroupements d'individus d'une part, et de parties de variables d'autre part. Nous appliquons cette méthodologie sur plusieurs jeux de donnée en la comparant aux résultats d'une analyse par correspondances multiples ACM, appliquée aux même données binarisées.

1 Introduction

Les méthodes d'analyse de données peuvent être regroupées en deux grandes catégories : l'analyse supervisée où l'objectif est de prédire une variable cible à partir de variables explicatives et l'analyse non-supervisée où l'objectif est de découvrir la structure sous-jacente des données en regroupant les individus dans des groupes homogènes (clustering). Apparue comme extension du clustering, la classification croisée (Good (1965); Hartigan (1975)), appelée aussi coclustering, est une technique non-supervisée dont l'objectif est d'effectuer une classification simultanée des individus et des variables d'un tableau de données. De nombreuses méthodes ont été développées pour effectuer de la classification croisée (par exemple : Bock (1979); Govaert (1983); Dhillon et al. (2003); Govaert et Nadif (2013)). Ces méthodes diffèrent principalement dans le type des données étudiées (continues, binaires ou de contingence), les hypothèses considérées, la méthode d'extraction utilisée et les résultats souhaités. En particulier, deux grandes familles de méthodes ont été largement étudiées : les méthodes de reconstruction de matrices où le problème est présenté sous forme d'approximation matricielle et les méthodes basées sur les modèles de mélange où les blocs sont définis par des variables