Un modèle Bayésien de co-clustering de données mixtes

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

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

Résumé. Nous proposons un modèle de co-clustering de données mixtes et un critère Bayésien de sélection du meilleur modèle. Le modèle infère automatiquement les discrétisations optimales de toutes les variables et effectue un co-clustering en minimisant un critère Bayésien de sélection de modèle. Un avantage de cette approche est qu'elle ne nécessite aucun paramètre utilisateur. De plus, le critère proposé mesure de façon exacte la qualité d'un modèle tout en étant régularisé. L'optimisation de ce critère permet donc d'améliorer continuellement les modèles trouvés sans pour autant sur-apprendre les données. Les expériences réalisées sur des données réelles montrent l'intérêt de cette approche pour l'analyse exploratoire des grandes bases de données.

1 Introduction

Dans un monde où les technologies d'acquisition de données sont en croissance rapide, l'analyse exploratoire des bases de données hétérogènes et de grandes tailles reste un domaine peu étudié. Une technique fondamentale de l'analyse non supervisée est celle du clustering, dont l'objectif est de découvrir la structure sous-jacente des données en regroupant les individus *similaires* dans des groupes homogènes. Cependant, dans de nombreux contextes d'analyse exploratoire de données, cette technique de regroupement d'objets reste insuffisante pour découvrir les motifs les plus pertinents. Le co-clustering (Hartigan, 1975), apparu comme extension du clustering, est une technique non-supervisée dont l'objectif est regrouper conjointement les deux dimensions de la même table de données, en profitant de l'interdépendance entre les deux entités (individus et variables) représentées par ces deux dimensions pour extraire la structure sous-jacente des données. Cette technique est la plus adaptée, par example, dans des contextes comme l'analyse des paniers de consommation où l'objectif est d'identifier les sous-ensembles de clients ayant tendance à acheter les mêmes de produits, plutôt que de grouper simplement les clients (ou les produits) en fonction des modèles d'achat/vente.

Dans la littérature, plusieurs approches de co-clustering ont été développées. En particulier, certains algorithmes de co-clustering proposent d'optimiser une fonction qui mesure l'écart entre la matrice de données et la matrice de co-clusters (Cheng et Church, 2000). D'autres techniques sont basées sur la théorie de l'information (Dhillon et al. (2003)), sur les modèles de mélange pour définir des modèles de blocs latents (Govaert et Nadif, 2008), sur l'estimation Bayésienne des paramètres (Shan et Banerjee (2008)), sur l'approximation matricielle (Lee et