Sélection Bayésienne de Modèles avec Prior Dépendant des Données

Marc Boullé *

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
marc.boulle@orange.com,
http://perso.rd.francetelecom.fr/boulle/

Résumé. Cet article analyse la consistance asymptotique des modèles en grille appliqués à l'estimation de densité jointe de deux variables catégorielles. Les modèles en grille considèrent un partitionnement des valeurs de chacune des variables, le produit Cartésien des partitions formant une grille dont les cellules permettent de résumer la table de contingence des deux variables. Le meilleur modèle de co-partitionnement est recherché au moyen d'une approche MAP (maximum a posteriori), présentant la particularité peu orthodoxe d'exploiter une famille de modèles et une distribution a priori de ces modèles qui dépendent des données. Ces modèles sont par nature des modèles de l'échantillon d'apprentissage, et non de la distribution sous-jacente. Nous démontrons la consistance de l'approche, qui se comporte comme un estimateur universel de densité jointe convergeant asymptotiquement vers la vraie distribution jointe.

1 Introduction

L'analyse de la corrélation entre deux variables catégorielles est un problème largement étudié en analyse des données. Dans le cas de variables ayant de nombreuses valeurs, le tableau de contingence des deux variables peut être résumé au moyen d'un tableau synthétique par groupement des lignes et des colonnes. Ce problème est traité par maximisation d'un critère de corrélation entre les variables (Ritschard et al., 2001), ou formulé en tant que problème de coclustering des lignes et des colonnes d'une matrice (Hartigan, 1972) avec application au groupement simultané des individus et des variables dans (Bock, 1979; Nadif et Govaert, 2005; Dhillon et al., 2003).

Dans ce papier, nous étudions l'approche MODL par modèles en grilles (Boullé, 2011a) dans le cas de deux variables catégorielles, qui a été appliquée dans le cas du coclustering des textes et mots d'un corpus (Poirier et al., 2008) ou de celui des noeuds sources et cibles d'un graphe de grande taille (Boullé, 2011b). La famille de modèles envisagée est un partitionnement bivarié des deux variables, la table de contingence résultante permettant de résumer la corrélation entre les variables. L'approche est régularisée au moyen d'une approche MAP qui permet de déterminer automatiquement la granularité du co-partitionnement. La validation expérimentale de la méthode sur des bases comportant des dizaines de milliers de valeurs et des millions d'individus a démontré d'excellentes performances, avec la détection fiable de motifs précis au moyen d'algorithmes en $O(N\sqrt{N}\log N)$ où N est le nombre d'individus.