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.

Cependant, cette approche exploite une distribution a priori des paramètres de modélisation qui dépend des données : cela pose d'évidents problèmes de consistance, l'approche modélisant l'échantillon d'apprentissage et non la distribution sous-jacente. Dans cet article, nous démontrons la consistance de l'approche, qui se comporte comme un estimateur universel de densité jointe convergeant asymptotiquement vers la vraie distribution jointe.

Le reste de l'article est organisé de la façon suivante. La partie 2 rappelle l'approche MODL dans le cas du résumé de deux variables catégorielles. La consistance asymptotique de l'approche est présentée dans la partie 3. Enfin, la partie 4 conclut cet article.

2 Approche MODL pour deux Variables Catégorielles

On rappelle ici l'approche MODL des modèles en grille dans le cas de deux variables catégorielles X et Y, dont on cherche à décrire conjointement les valeurs. On introduit en définition 1 une famille de modèles où chaque variable est partitionnée en groupes de valeurs (modalités). Les individus sont distribués sur l'ensemble des cellules de la grille bidimensionnelle résultant du produit Cartésien des partitions univariées ainsi définies. Cette distribution étant spécifiée, on en déduit par sommation sur les cellules la distribution des individus sur les groupes de valeurs pour chaque variable. Il ne reste qu'à spécifier localement à chaque groupe la distribution des individus sur les valeurs du groupe pour obtenir une description complète de la distribution des individus sur les valeurs des deux variables conjointement.

Définition 1. *Un modèle de groupement de valeurs bivarié est défini par :*
- *un nombre de groupes pour chaque variable,*
- *la partition de chaque variable en groupes de valeurs,*
- *la distribution des individus sur les cellules de la grille de données ainsi définie,*
- *la distribution des individus de chaque groupe sur les valeurs du groupe, par variable.*

Notations 1.
- N *: nombre d'individus de l'échantillon*
- V, W *: nombre de valeurs pour chaque variable (connu)*
- I, J *: nombre de groupes pour chaque variable (inconnu)*
- $G = IJ$ *: nombre de cellules de la grille du modèle*
- $i(v), j(w)$ *: index du groupe auquel est rattachée la valeur* v *(resp.* w*)*
- $m_{i.}, m_{.j}$ *: nombre de valeurs du groupe* i *(resp.* j*)*
- $n_{v.}, n_{.w}$ *: nombre d'individus pour la valeur* v *(resp.* w*)*
- n_{vw} *: nombre d'individus pour la paire de valeurs* (v, w)
- $N_{i.}, N_{.j}$ *: nombre d'individus du groupe* i *(resp.* j*)*
- N_{ij} *: nombre d'individus de la cellule* (i, i) *de la grille*

Un modèle de groupement de valeurs bivarié est entièrement caractérisé par le choix des paramètres de partition des valeurs en groupes I, J, $\{i(v)\}_v$, $\{j(w)\}_w$, des paramètres de distribution des individus sur les cellules de la grille $\{N_{ij}\}_{i,j}$, et des paramètres de distribution des individus des groupes sur les valeurs des variables $\{n_{v.}\}_v$, $\{n_{.w}\}_w$. Les nombres de valeurs par groupe sont déduits du choix des partitions des valeurs en groupes, et les effectifs des groupes par comptage des effectifs des cellules de la grille. Afin de rechercher le meilleur modèle, on applique une approche MAP visant à maximiser la probabilité

$P(M|D) = P(M)P(D|M)/P(D)$ du modèle sachant les données. À cet effet, on introduit une distribution a priori sur les paramètres des modèles, exploitant la hiérarchie des paramètres de modélisation, uniforme à chaque étage de cette hiérarchie. Les paramètres de modélisation dépendent des caractéristiques de l'échantillon (N, V, W) et sont à valeurs discrètes. Ainsi, l'estimation des probabilités pour une distribution uniforme se fait par comptage à chaque étage de la hiérarchie. Par exemple, pour une grille de taille donnée (I, J), le nombre de façons de distribuer N individus sur les $G = IJ$ cellules de la grille est de $\binom{N+G-1}{G-1}$, ce qui donne une probabilité de $1/\binom{N+G-1}{G-1}$ par jeu de paramètres discrets de multinomiale.

En utilisant la définition formelle des modèles et leur distribution a priori hiérarchique, la formule de Bayes permet de calculer de manière exacte la probabilité d'un modèle connaissant les données, ce qui conduit au théorème 1.

Théorème 1. *Le log négatif de la probabilité a posteriori d'un modèle de groupement de valeurs bivarié M suivant un a priori hiérarchique uniforme est donné par la formule suivante :*

$$\begin{aligned} c(M) =& \log V + \log W + \log B(V, I) + \log B(W, J) \\ &+ \log \binom{N+G-1}{G-1} + \sum_{i=1}^{I} \log \binom{N_{i.}+m_{i.}-1}{N_{i.}-1} + \sum_{j=1}^{J} \log \binom{N_{.j}+m_{.j}-1}{N_{.j}-1} \\ &+ \log N! - \sum_{i=1}^{I}\sum_{j=1}^{J} \log N_{ij}! \\ &+ \sum_{i=1}^{I} \log N_{i.}! + \sum_{j=1}^{J} \log N_{.j}! - \sum_{v=1}^{V} \log n_{v.}! - \sum_{w=1}^{W} \log n_{.w}! \end{aligned} \tag{1}$$

$B(V, I)$ est le nombre de répartitions de V valeurs explicatives en I groupes (éventuellement vides). Pour $I = V$, $B(V, I)$ correspond au nombre de Bell. Dans le cas général, $B(V, I)$ peut s'écrire comme une somme de nombres de Stirling de deuxième espèce.

La première ligne de la formule (1) regroupe les termes d'a priori correspondant au choix des nombres de groupes I et J et à la spécification de la partition de chaque variable en groupes de valeurs. La deuxième ligne représente la spécification de la distribution multinomiale des N individus de l'échantillon sur les G cellules de la grille, suivi de la spécification de la distribution des individus de chaque groupe sur les valeurs du groupe. La troisième ligne représente la vraisemblance de la distribution des individus dans les cellules de la grille, au moyen d'un terme du multinôme. La dernière ligne correspond à la vraisemblance des valeurs localement à chaque groupe pour chacune des variables.

3 Consistance Asymptotique de l'Approche

Dans cette partie, les modèles présentés en partie 2 sont interprétés comme des estimateurs de densité jointe, puis la convergence asymptotique de l'approche MODL vers la vraie densité est présentée.

Pour deux variables catégorielles X et Y ayant V et W valeurs, la densité jointe est entièrement définie au moyen de VW paramètres de probabilité $p(v, w) = p(X = v, Y = w)$. Pour un modèle en grille M défini selon la définition 1, on introduit les notations suivantes :

- $p_{ij} = \frac{N_{ij}}{N}$: probabilité d'appartenir à la cellule (i, j) de la grille,
- $p_{v,i.} = \frac{n_{v.}}{N_{i.}}$: probabilité de la valeur v du groupe i de X,
- $p_{w,.j} = \frac{n_{.w}}{N_{.j}}$: probabilité de la valeurs w du groupe j de Y.

Les modèles en grille sont des estimateurs de densité jointe constants par cellule par rapport aux cellules de la grille bivariée, contraints par les probabilités marginales de X et Y. En faisant l'hypothèse de l'indépendance des variables localement à chaque cellule, on obtient l'estimation de densité jointe ci-dessous :

$$p_{vw} = p_{ij} p_{v,i.} p_{w,.j} = \frac{N_{ij}}{N} \frac{n_{v.}}{N_{i.}} \frac{n_{.w}}{N_{.j}}, \tag{2}$$

où (i, j) est la cellule de la grille contenant les valeurs v de X et w de Y.

Les paramètres des modèles en grille sont discrets, et leur distribution a priori dépend de données : c'est l'échantillon de taille finie N avec V et W valeurs qui est directement modélisé, et non la distribution bivariée sous-jacente. Nous étudions maintenant si cette approche de modélisation converge asymptotiquement vers la vraie densité jointe sous-jacente quand le nombre d'individus tend vers l'infini. Rappelons d'abord quelques notions de théorie de l'information. L'entropie de Shannon $H(X)$ (Shannon, 1948) d'une variable discrète X est définie par :

$$H(X) = -\sum_{x \in X} p(x) \log p(x). \tag{3}$$

L'information mutuelle entre deux variables est une mesure de leur dépendance mutuelle (Cover et Thomas, 1991), qui s'annule si et seulement si les deux variables sont indépendantes. Pour deux variables discrètes X et Y, l'information mutuelle est définie par :

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{4}$$

Pour un modèle de groupement de valeurs bivarié M, considérons les variables groupées X_M and Y_M ayant I et J valeurs. Le théorème 2 présente une approximation asymptotique du critère $c(M)$ introduit dans la formule 1.

Théorème 2. *Le critère d'évaluation MODL (formule 1) pour un modèle de groupement de valeurs bivarié M est asymptotiquement égale à N fois la somme de l'entropie des variables diminué de l'information mutuelle entre les variables groupées.*

$$c(M) = N\left(H(X) + H(Y) - I(X_M; Y_M)\right) + O(\log N). \tag{5}$$

Le critère étant à minimiser, cela signifie que la méthode vise à sélectionner le modèle qui maximise l'information mutuelle entre X et Y. Comme l'information mutuelle entre deux variables correspond à la divergence de Kullback-Leibler (Kullback, 1959) entre la distribution jointe des deux variables et leur distribution jointe en cas d'indépendance (produit des distributions marginales), cela signifie que le modèle le plus probable (MAP) sélectionné selon l'approche MODL vise à maximiser le "contraste" entre les variables, en s'éloignant autant que possible de leur distribution jointe en cas d'indépendance.

Nous présentons maintenant avec le théorème 3 le résultat central de l'article, qui montre que l'approche MODL converge asymptotiquement vers l'estimation de la vraie densité jointe entre X et Y. Bien que la technique de modélisation soit dépendante des données (en regard de l'espace des modèles et du choix de la distribution a priori des paramètres de modélisation) et vise à modéliser l'échantillon de données directement au moyen d'une distribution discrète des individus sur les valeurs des variables, ce théorème démontre la consistance de l'approche, qui estime asymptotiquement la vraie distribution jointe de probabilité, à valeurs réelles.

Théorème 3. *Le modèle de groupement de valeurs bivarié M sélectionné selon l'approche MODL converge asymptotiquement vers la vraie distribution jointe des deux variables, et le critère pour le meilleur modèle M_{Best} converge vers N fois l'entropie jointe des deux variables.*

$$\lim_{m \to \infty} \frac{c(M_{Best})}{N} = H(X, Y). \tag{6}$$

En corollaire du théorème 3, le théorème 4 indique que l'approche MODL permet d'estimer l'information mutuelle entre les variables.

Théorème 4. *Le modèle de groupement de valeurs bivarié M sélectionné selon l'approche MODL converge asymptotiquement vers la vraie distribution jointe des deux variables, et le critère pour le modèle nul $M_\emptyset$ diminué du critère pour le meilleur modèle M_{Best} converge vers N fois l'information mutuelle des deux variables.*

$$\lim_{m \to \infty} \frac{c(M_\emptyset) - c(M_{Best})}{N} = I(X; Y). \tag{7}$$

Le théorème 4 montre qu'en cas de deux variables indépendantes, le meilleur modèle sera asymptotiquement le modèle nul (qui représente effectivement le cas de deux variables indépendantes). Comme l'approche MODL est régularisée, avec des termes d'a priori dans le critère $c(M)$ qui augmentent avec la granularité du modèle, on s'attend à ce que l'approche sélectionne le modèle nul y compris dans le cas non asymptotique. La robustesse de l'approche, observée lors de nombreuses expérimentations sur des jeux de données aléatoires (Boullé, 2011a,b), reçoit ici une explication théorique.

4 Conclusion

Les modèles en grilles considèrent le partitionnement d'un ensemble de variables, en intervalles dans le cas numérique et en groupes de valeurs dans le cas catégoriel, le produit Cartésien de ces partitions univariées formant une grille de cellules permettant d'estimer la densité jointe entre les variables. L'approche MODL est une approche MAP de sélection de modèles exploitant à la fois une famille de modèles et une distribution a priori des paramètres dépendant des données. Selon cette approche peu orthodoxe, c'est l'échantillon directement qui est modélisé, et non la distribution sous-jacente. Dans cet article, nous avons étudié cette approche dans le cas de deux variables catégorielles et démontré qu'elle se comporte comme un estimateur universel de densité jointe convergeant asymptotiquement vers la vraie distribution. Dans des travaux futurs, nous envisageons d'étendre ces résultats de consistance aux grilles comportant un nombre quelconque de variables, numériques ou catégorielles.

Références

Bock, H. (1979). Simultaneous clustering of objects and variables. In E. Diday (Ed.), *Analyse des Données et Informatique*, pp. 187–203. INRIA.

Boullé, M. (2011a). *Data grid models for preparation and modeling in supervised learning*, pp. 99–130. Microtome Publishing.

Boullé, M. (2011b). Estimation de la densité d'arcs dans les graphes de grande taille : une alternative à la détection de clusters. In *Extraction et gestion des connaissances (EGC'2011)*, pp. 353–364.

Cover, T. et J. Thomas (1991). *Elements of information theory*. New York, NY, USA : Wiley-Interscience.

Dhillon, I. S., S. Mallela, et D. S. Modha (2003). Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, pp. 89–98.

Hartigan, J. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association 67*(337), 123–129.

Kullback, S. (1959). *Information theory and statistics*. New York : John Wiley and Sons. republished by Dover, 1968.

Nadif, M. et G. Govaert (2005). Block clustering of contingency table and mixture model. In *Advances in Intelligent Data Analysis VI*, Volume 3646 of *LNCS*, pp. 249–259. Springer.

Poirier, D., C. Bothorel, et M. Boullé (2008). Analyse exploratoire d'opinions cinématographiques : co-clustering de corpus textuels communautaires. In *Extraction et gestion des connaissances (EGC'2008)*, pp. 565–576.

Ritschard, G., D. A. Zighed, et N. Nicoloyannis (2001). Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé. *Mathématiques et Sciences Humaines 154-155*, 81–98.

Shannon, C. (1948). A mathematical theory of communication. Technical Report 27, Bell systems technical journal.

Summary

This paper studies the asymptotic consistency of the data grid models applied to the joint density estimation of two categorical variables. The data grid models consider the grouping of the values of each variable. The Cartesian product of these partitions forms a grid whose cells provide a summary of the contingency table of the two variables. The best bivariate grouping model is searched by the mean of a MAP (maximum a posteriori) approach, with the heretic property of exploiting both a model family and a prior distribution that are data dependent. These models are in essence models of the data sample, not of the underlying distribution. We demonstrate the consistency of the approach, which behaves as a universal estimator of joint density that asymptotically converges towards the true joint distribution.