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

latentes à estimer (voir Brault et Lomet (2015) pour une revue des méthodes de classification jointe). Dans ce dernier cas, des mélanges de gaussiennes sont souvent utilisés pour modéliser les données continues et des mélanges de Bernoulli pour modéliser les données binaires.

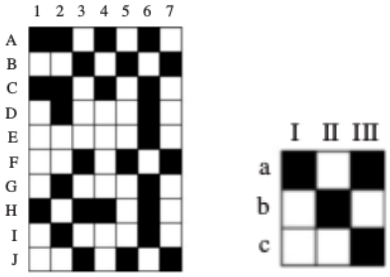

FIG. 1: Exemple de classification croisée, avec à gauche les données binaires initiales, et à droite après coclustering des lignes et colonnes.

La figure 1 montre l'exemple d'un tableau de données binaires représentant $n = 10$ individus et $p = 7$ variables binaires (Govaert et Nadif (2008)) et le tableau binaire résumant le résultat d'une classification croisée en $3 \times 3 = 9$ blocs binaires, résumé qui permet de visualiser plus simplement les principales associations.

Les méthodes de coclustering s'appliquent naturellement à des données de même nature, avec des variables devant être toutes binaires, numériques ou catégorielles. Nous proposons ici d'étendre ces méthodes d'analyse exploratoire selon 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. Le nombre de parties étant fixé, nous ne souhaitons pas imposer de paramètres supplémentaires, tels que le nombre de groupes d'individus ou de parties de variable. Pour ce faire, nous utilisons l'approche MODL (Boullé (2011)) de coclustering pour sa nature non paramétrique, son efficacité pour découvrir les structures de corrélation et sa capacité de passage à l'échelle.

Étant intéressé par l'étude des variables de types mixtes, nous comparons notre méthodologie à la méthode dérivée de l'analyse factorielle la plus utilisée en présence de données qualitatives : l'analyse par correspondances multiples ACM. En effet, l'ACM est une méthode permettant d'analyser les corrélations entre les variables qualitatives tout en réalisant une typologie des individus. Elle permet ainsi de traiter l'étude des individus et des variables comme deux problèmes complémentaires et les résout en dualité, les résultats sur les individus pouvant s'interpréter sur les variables et inversement. Ces objectifs sont cohérents avec ceux de la classification croisée selon notre approche, d'où l'intérêt de cette comparaison.

Le reste de cet article est organisé comme suit. Dans la section 2, nous rappelons l'approche de coclustering MODL, puis dans la section 3 nous présentons la méthodologie de coclustering entre individus et variables de types mixtes. Dans la section 4, nous résumons le principe et déroulement de l'analyse par correspondances multiples ACM. La section 5 présente les

résultats expérimentaux ainsi qu'une analyse comparative, et la section 6 les conclusions et perspectives.

2 Coclustering MODL de deux variables catégorielles

On résume ici l'approche MODL (Boullé (2011)) 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 2.1 une famille de modèles d'estimation de densité jointe entre les variables, sur la base d'une partition des valeurs de chaque variable en groupes de valeurs.

Définition 2.1. *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.*

Soient :
- 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
- $m_{i.}, m_{.j}$: 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

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 connaissant les données. A 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.

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 2.1.

Théorème 2.1. *Un modèle d'estimation de densité par grille suivant un a priori hiérarchique est optimal au sens de Bayes si son évaluation par la formule suivante est minimale sur l'ensemble de tous les modèles (Boullé (2011)) :*

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

où $B(V, I)$ est le nombre de répartitions des V valeurs en I groupes ce qui peut s'écrire comme une somme des nombres de Stirling de deuxième espèce : $B(V, I) = \sum_{i=1}^{I} S(V, i)$.

L'estimation de la densité jointe de deux variables catégorielles selon un a priori hiérarchique sur les paramètres est implémentée dans le logiciel Khiops (Boullé (2008)). Nous utilisons ce logiciel dans les expérimentations effectuées en section 5.

3 Coclustering entre individus et variables de type mixte

Nous présentons ici notre approche, qui consiste à binariser les variables numériques et catégorielles avant d'appliquer une méthode de coclustering individus x variables dans le cas de variables binaires.

3.1 Prétraitement des données

Dans une première étape, les variables sont binarisées selon un paramètre utilisateur k, représentant le nombre maximal de parties par variable. Dans le cas de variables numériques, ces parties sont obtenus selon une discrétisation non supervisée en k intervalles de fréquence égale. Dans le cas de variables catégorielles, les $k-1$ valeurs les plus fréquentes définissent les premiers parties, le dernier accueillant toutes les autres valeurs.

Ce paramètre k définit la granularité maximale à laquelle se fera l'analyse. Son choix repose sur un compromis entre finesse de l'analyse, temps de calcul du coclustering effectué en deuxième étape, et interprétabilité des résultats. Il est à noter que ce paramètre de granularité initiale des données est moins contraignant que les paramètres de type nombre de clusters d'individus ou de variables habituellement utilisés dans la plupart des méthodes de coclustering. Dans les expériences, nous utiliserons k=5 et $k = 10$. Pour la base Iris par exemple, le résultat de la binarisation des variables en 5 parties est illustré dans le tableau 1.

SepalLength	SepalWidth	PetalLength	PetalWidth	Class
$]-\infty; 5.05]$	$]-\infty; 2.75]$	$]-\infty; 1.55]$	$]-\infty; 0.25]$	Iris-setosa
$]5.05; 5.65]$	$]2.75; 3.05]$	$]1.55; 3.95]$	$]0.25; 1.15]$	Iris-versicolor
$]5.65; 6.15]$	$]3.05; 3.15]$	$]3.95; 4.65]$	$]1.15; 1.55]$	Iris-virginica
$]6.15; 6.55]$	$]3.15; 3.45]$	$]4.65; 5.35]$	$]1.55; 1.95]$	
$]6.55; +\infty[$	$]3.45; +\infty[$	$]5.35; +\infty[$	$]1.95; +\infty[$	

TAB. 1: les resultats de discretisation pour $k = 5$

3.2 Transformation des données en deux variables

L'approche MODL (Boullé (2011)) résumée en section 2 est intéressante pour son absence de paramètre utilisateur, son efficacité pour découvrir les structures de corrélation et sa capacité de passage à l'échelle. Bien que dédiée à l'estimation de densité jointe entre deux variables, elle a été appliquée dans le cas de coclustering entre individus et variables binaires, par exemple dans le cas d'un corpus de textes de grande taille, où chaque texte est décrit par une dizaine de milliers de variables binaires représentant l'utilisation d'un mot. Pour ce cas, le corpus de texte a été préalablement transformé en une représentation en deux variables, *IdText* et *IdMot*.

IdInstance	*IdVarPart*
$I1$	*SepalLength*$]5.05; 5.65]$
$I1$	*SepalWidth*$]3.45; +\infty[$
$I1$	*PetalLength*$] - \infty; 1.55]$
$I1$	*PetalWidth*$] - \infty; 0.25]$
$I1$	*Class{Iris-setosa}*
$I2$	*SepalLength* $] - \infty; 5.05]$
$I2$	*SepalWidth*$]2.75; 3.05]$
$I2$	*PetalLength*$] - \infty; 1.55]$
$I2$	*PetalWidth*$] - \infty; 0.25]$
$I2$	*Class{Iris-setosa}*

TAB. 2: Les 10 premieres instances de la base Iris binarisée

De la même façon, on transforme ici le jeu de donnée binarisé en deux variables, *IdInstance* et *IdVarPart*, en créant pour chaque individu initial un enregistrement par variable, mémorisant le lien entre l'individu et sa partie de variable. L'ensemble des n individus initiaux représentés par m variables est ainsi transformé en un nouveau jeu de données de taille $N = nm$ ayant deux variables catégorielles, la première comportant $V = N$ valeurs et la seconde (au plus) $W = m \times k$ valeurs. Dans la base Iris par exemple, cette étape résulte en deux colonnes de 750 instances. Le tableau 2 montre les dix premiers instances.

3.3 Coclustering et interprétation des résultats

Les données étant représentées sous forme de deux variables, la méthode MODL est appliquée pour rechercher un modèle d'estimation de densité jointe entre ces deux variables. Dans le coclustering résultat, les individus de la base initiale (valeurs de la variable *IdInstance*) sont regroupés s'ils sont distribués de façon similaire sur les groupes de parties de variables (valeurs de la variable *IdVarPart*), et réciproquement.

4 L'analyse des correspondances multiples ACM

L'analyse factorielle est un ensemble de méthodes qui s'appliquent aux tableaux de données dont les lignes représentent les individus et les colonnes représentent les variables (de type quelconque). Les questions qui se posent en analyse factorielle sont celles de ressemblance ou dissimilarité entre des groupes d'individus (problème étudié dans la classification non supervisée) et des niveaux de liaisons (corrélations) entre les variables. L'analyse de correspondances multiples permet d'analyser les corrélations entre les variables qualitatives, de transformer les variables qualitatives en quantitatives et de réaliser une typologie d'individus et des variables de manière complémentaire.

4.1 L'analyse des correspondances multiples en pratique

Considérons le tableau de données individus×variables $\mathbf{x} = (x_{ij}, i \in I, j \in J)$ où I est un ensemble de n objets étudiés et J l'ensemble des p variables qualitatives (de m_j modalités chacune) caractérisant les objets (représentés, respectivement, par les lignes et les colonnes de la matrice $\mathbf{x}$). Les opérations mathématiques n'ayant pas de sens sur les variables catégorielles, l'ACM passe par la représentation de ces mêmes données sous forme d'un Tableau Disjonctif Complet (TDC) ; une juxtaposition des p tableaux d'indicatrices des variables où les lignes représentent les individus et les colonnes représentent les modalités des variables. Ce tableau peut être considéré comme un tableau de contingence entre les individus et les modalités des variables. Pour un TDC donné, noté ici T, la somme des éléments de chaque ligne de T est égale à p le nombre de variables, la somme des éléments d'une colonne s de T donne l'effectif marginal de la modalité correspondante (noté n_s), la somme des colonnes de chaque tableau d'indicatrices est égale au vecteur 1, la somme de tous les éléments de T est (np), la matrice des poids des lignes de T est $r = \frac{1}{n} I$, la matrice des poids des colonnes de T est la matrice diagonale $D = diag(D_1, D_2, \ldots, D_p)$ où chaque D_j correspond à la matrice diagonale contenant les fréquences marginales des modalités de la j^{eme} variable.

4.2 Principaux résultats de l'ACM

Les coordonnées des modalités sur les axes factoriels sont les vecteurs propres de la solution $\frac{1}{p}\mathbf{D}^{-1}\mathbf{T}^t\mathbf{T}$ de l'équation :

$$\frac{1}{p}\mathbf{D}^{-1}\mathbf{T}^t\mathbf{T}\mathbf{a} = \mu\mathbf{a}$$

Les coordonnées des individus sur les axes factoriels sont les vecteurs propres de $\frac{1}{p}\mathbf{T}\mathbf{D}^{-1}\mathbf{T}^t$, solution de l'équation :

$$\frac{1}{p}\mathbf{T}\mathbf{D}^{-1}\mathbf{T}^t\mathbf{z} = \mu\mathbf{z}$$

On déduit (Saporta (2006)) les formules de transition : $z = \frac{1}{\sqrt{\mu}}\frac{1}{p}\mathbf{T}\mathbf{a}$ et $\mathbf{a} = \frac{1}{\sqrt{\mu}}\mathbf{D}^{-1}\mathbf{T}^t\mathbf{z}$

Notons que :

— l'inertie totale du nuage étudié vaut $(\frac{m}{p} - 1)$

— l'inertie des m_j modalités de la variable V_j vaut $\frac{1}{p}(m_j - 1)$. Cette inertie, étant liée directement au nombre de modalités de la variable V_j, il est préférable d'exiger des nombres de modalités égaux pour toutes les variables actives, d'où l'intérêt du prétraitement (section 3.1).

— les contributions de l'individu i et de la modalité s sur un axe factoriel h sont données par :

$$Ctr_h(i) = \frac{1}{n}\frac{z_{ih}^2}{\mu_h} \text{ et } Ctr_h(s) = \frac{n_s}{np}\frac{a_{ih}^2}{\mu_h}$$

— la contribution d'une variable à l'inertie d'un facteur permettant de mesurer la liaison entre cette variable et ce facteur est la somme des contributions de toutes ses modalités.

L'ACM permet d'analyser simultanément des variables qualitatives et quantitatives. Pour cela, nous suivons l'approche classique de décomposer la plage de valeurs des chaque variable quantitative en plusieurs intervalles (classes).

5 Expérimentation

Nous comparons les méthodes de coclustering (section 3) et ACM (section 4) sur la base Iris pour des raisons didactiques, puis nous évaluons notre approche sur la base Adult (Lichman (2013)) pour évaluer le passage à l'échelle.

5.1 Comparaison des méthodes sur la base Iris

La base Iris comporte $n = 150$ individus et $m = 5$ variables, quatre numériques et une catégorielle.

5.1.1 Coclustering

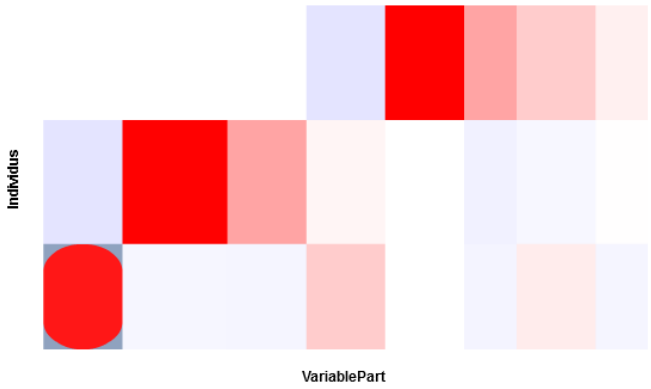

FIG. 2: Coclustering pour la base Iris

Après avoir binarisé les données de la base Iris selon une granularité en $k = 5$ parties et appliqué la méthode de coclustering MODL, la grille optimale obtenue consiste en 3 clusters d'individus et 8 clusters de parties de variables. La figure 2 présente cette grille, avec en ligne les clusters d'individus, en colonne les clusters de parties de variables. L'information mutuelle entre les deux dimensions est visualisée dans les cellules, en rouge dans le cas de sur-représentation des instances dans la cellule par rapport à l'hypothèse où les deux dimensions seraient indépendantes, en bleu en cas de sous-représentation. Les trois clusters d'individus visibles sur la figure 2 peuvent être caractérisés d'une part par les types de fleurs dont ils sont composés, d'autre part par les parties de variables les plus représentées pour ces clusters (cellule rouge la plus sur-représentée de chaque ligne) :

— en haut : cluster de 50 fleurs, toutes de la classe *Iris-setosa*, et caractérisé par les parties de variables *Class{Iris-setosa}*, *PetalLength*$]-inf;1.55]$ et *PetalWidth*$]-inf;0.25]$,

— au milieu : cluster de 54 fleurs, dont 50 de la classe *Iris-virginica*, et caractérisé par les parties de variables données par : *Class{Iris-virginica}*, *PetalLength*$]5.35;+inf[$, *PetalWidth*$]1.95;+inf[$ et *PetalWidth*$]1.55;1.95]$,

— en bas : cluster de 46 fleurs, toutes de la classe *Iris-versicolor*, et caractérisé par les parties de variables données par : *Class{Iris-versicolor}*, *PetalLength*$]3.95; 4.65]$ et *PetalWidth*$]1.15; 1.55]$.

De façon intéressante, les trois clusters d'individus sont facilement interprétables : il s'agit de haut en bas des *petites*, *grandes* et *moyennes* fleurs. Ces clusters sont expliqués principalement par trois clusters de parties de variables, faisant toutes intervenir les variables *Class*, *PetalLength* et *PetalWidth*.

De façon duale, en regardant les clusters de parties de variables, on trouve deux clusters de parties de variables (la quatrième et la huitième colonne) peu informatifs (les deux colonnes les moins contrastées claires), et basés essentiellement sur la variable *SepalWidth* :

— quatrième colonne : contient les parties *SepalWidth*$]-\infty; 2.75]$, *SepalWidth*$]2.75; 3.05]$, *SepalLength*$]5.65; 6.15]$,

— huitième colonne : contient les parties *SepalWidth*$]3.05; 3.15]$, *SepalWidth*$]3.15; 3.45]$.

Les faibles valeurs de *SepalWidth* (quatrième colonne) sont légèrement sur-représentées pour les clusters d'individus associés aux classes *Iris-versicolor* et *Iris-virginica*, alors que les valeurs intermédiaires (huitième colonne) sont légèrement sur-représentées pour le cluster d'individus associé à la classe *Iris-versicolor*.

5.1.2 Analyse ACM

L'analyse sur les bases Iris est effectuée sur la base de la même binarisation des variables que précédemment.

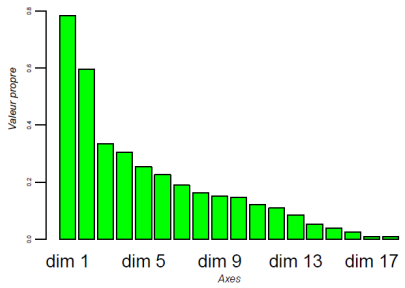

FIG. 3: Histogramme des valeurs propres pour l'analyse ACM de Iris

La distribution des valeurs propres (Figure 3) indique que les deux premiers axes factoriels capturent suffisamment de variabilité pour limiter l'analyse au premier plan factoriel.

La comparaison de la projection des variables (Figure 4 droite) et de la projection des individus (Figure 4 gauche) sur ce premier plan factoriel fait apparaître une nette corrélation de certaines variables entre elles :

— en haut à gauche du premier plan factoriel, Iris-virginica est corrélé avec les fortes valeurs de PetalLength (supérieures à 4.65), les fortes valeurs de PetalWidth (supérieures à 1.55) et les fortes valeurs de SepalLength (supérieures à 6.15)

— à droite du premier plan factoriel, Iris-setosa est fortement corrélé avec les faibles valeurs de PetalLength (inférieures à 3.95), les faibles valeurs de PetalWidth (inférieures à 1.15) et les faibles valeurs de SepalLength (inférieures à 5.05)

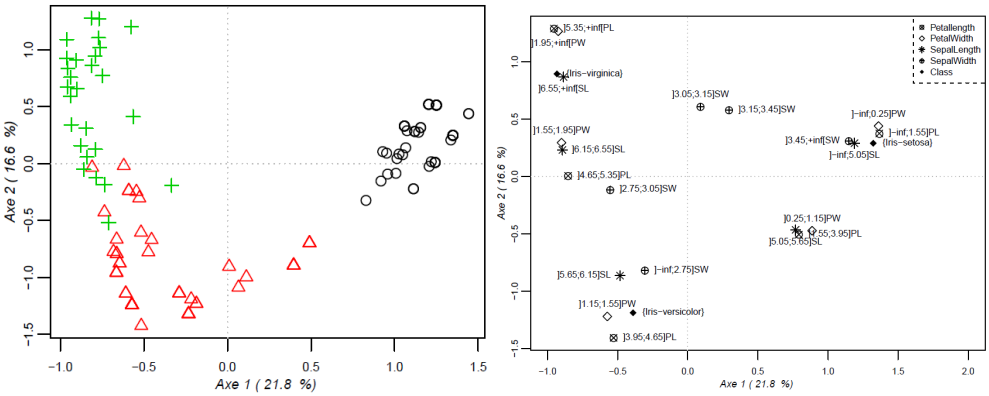

FIG. 4: Projection des individus et des variables de Iris sur les deux premiers axes

— en bas à gauche du premier plan factoriel, Iris-versicolor est corrélé avec les valeurs intermédiaires de PetalLength, PetalWidth et SepalWidth

La projection des individus (Figure 4 gauche) montre l'existence d'une zone de mélange entre Iris-virginica et Iris-versicolor. Ces résultats sont identiques à ceux qu'on a pu déduire de l'analyse par coclustering.

Les variables issues de SepalWidth montrent une plus faible corrélation avec les autres et sous-tendent moins le premier plan factoriel : les faibles valeurs (inférieures à 3.05 sont associées à la zone de mélange Iris-virginica et Iris-versicolor, les valeurs intermédiaires (entre 3.05 et 3.45) se projettent entre les nuages Iris-virginica et Iris-setosa (et sont donc présentes dans les deux populations). Ces résultats sont également en accord avec ceux déduits du coclustering (voir ci-dessus l'interprétation des colonnes 4 et 8 du coclustering).

Finalement, sur cet exemple didactique où l'interprétation de l'ACM peut se faire par simple inspection du premier plan factoriel, on fait apparaître un très bon accord entre ACM et coclustering.

5.2 Coclustering de la base Adult

La base Adult comporte $n = 48842$ individus représentés par $m = 15$ variables, 6 numériques et 9 catégorielles.

Après binarisation en $k = 10$ parties et transformation selon la méthodologie présentée en section 3, on obtient un jeu de données comportant $N \approx 750000$ lignes et deux colonnes : la variable *IdInstance* comportant environ $n \approx 50000$ valeurs (les individus) et la variable *IdVarPart* comportant $m \times k \approx 150$ valeurs (les parties de variable). L'algorithme de coclustering est un algorithme *anytime* qui publie régulièrement son indice de qualité (taux de compression atteint par le modèle). Pour la base Adult, le coclustering nécessite environ 4 mn de temps de calcul pour un premier résultat de qualité (le taux de compression ne varie plus de façon significative), et nous avons poursuivi l'optimisation pendant environ une heure pour une amélioration d'environ 5% de la log vraisemblance du modèle. Le résultat obtenu est très fin, avec 34 clusters d'individus et 62 clusters de parties de variables. Dans le cas de l'analyse exploratoire, cette finesse des résultats nuit à l'interprétabilité. Il est ici possible de simplifier

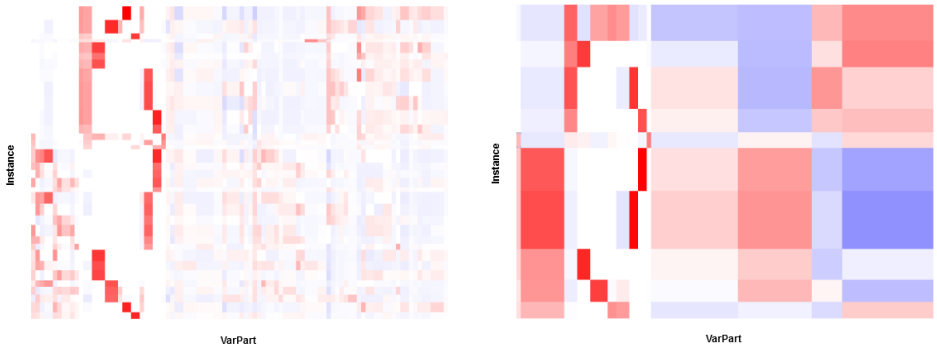

FIG. 5: Coclustering pour la base Adult, avec à gauche 100% et à droite 70% de l'information

le résultat d'analyse en fusionnant itérativement les lignes et les colonnes du coclustering le plus fin, jusqu'à garder un pourcentage donné de l'information initiale. La figure 5 présente ainsi les résultats du coclustering, avec à gauche un coclustering en 34 x 62 cellules contenant toute l'information, et à droite une version simplifiée en 10 x 14 cellules conservant 70% de l'information.

Le premier niveau de structuration dans les données apparait clairement comme une découpe en deux parties des clusters d'individus, visibles respectivement sur les moitiés hautes et basses des matrices de coclustering représentées sur la figure 5. Les clusters d'individus du haut sont principalement des hommes mariés plutôt aisés, avec une sur-représentation sur les clusters de parties incluant *sex{Male}*, *relationship{Husband}*, *relationship{Married...}*, *class{More}*, *age*$]45.5; 51.5]$, *age*$]51.5; 58.5]$, *hoursPerWeek*$]48.5; 55.5]$, *hoursPerWeek*$]55.5; +\infty[$. Les cluster d'individus du bas sont principalement des femmes ou des hommes non mariés plutôt pauvres, avec une sur-représentation sur les clusters de parties incluant *class{Less}*, *sex{Female}*, *maritalStatus{Never-married}*, *maritalStatus{Divorced}*, *relationship{Own-child}*, *relationship{Not-in-family}*, *relationship{Unmarried}*.

Dans la figure de gauche, le cluster d'individus le plus contrasté, donc le plus informatif, est sur la première ligne en partant du haut. Il s'interprète aisément en inspectant les clusters de partie de variable les plus sur-représentés sur cette ligne :

— *relationship{Husband}*, *relationship{Married...}*,
— *educationNum*$]13.5; +\infty[$, *education{Masters}*,
— *education{Prof-school}*,
— *sex{Male}*,
— *class{more}*,
— *occupation{Prof-specialty}*,
— *age*$]45.5; 51.5]$, *age*$]51.5; 58.5]$,
— *hoursPerWeek*$]48.5; 55.5]$, *hoursPerWeek*$]55.5; +\infty[$.

Il s'agit donc d'un cluster d'environ 2000 individus, avec principalement des hommes mariés ayant fait des études longues, travaillant dans l'enseignement en fin de carrière, travaillant beaucoup et gagnant bien leur vie.

Dans la figure de droite, les clusters de variables les plus contrastés, donc les plus informatifs, sont portés par les colonnes 4 à 9, et ne contiennent que des parties de variables *education*

et *educationNum*, qui sont les plus structurantes pour ce jeu de données.

— *educationNum*$]11.5; 13.5]$, *education{Assoc-acdm}*, *education{Bachelors}*,
— *educationNum*$]-\infty; 7.5]$, *education{10th}*, *education{11th}*, *education{7th-8th}*,
— *educationNum*$]13.5; +\infty[$, *education{Masters}*,
— *educationNum*$]10.5; 11.5]$, *education{Assoc-voc}*, *education{Prof-school}*,
— *educationNum*$]7.5; 9.5]$, *education{HS-grad}*,
— *educationNum*$]9.5; 10.5]$, *education{Some-college}*.

Les variables *education* et *educationNum* sont respectivement catégorielle et numérique et très corrélées entre elles : leurs clusters de parties présentés ci-dessus apparaissent particulièrement cohérents.

Un apport important de notre méthodologique, par rapport à l'ACM, réside dans sa facilité d'application et dans l'interprétabilité directe des résultats. Avec une ACM sur une base de taille importante (comme Adult et au-delà), la projection des individus et des variables sur le premier et même le deuxième plan factoriel ne permet souvent pas de distinguer des zones denses. De plus, il est souvent nécessaire de sélectionner un grand nombre d'axes. Sur la base Adult par exemple, les règles de sélection du nombre d'axes à interpréter indiquent qu'il faut choisir au moins 8 ou 9 axes ce qui fait qu'un traitement a posteriori est nécessaire pour pouvoir distinguer des groupes comme par exemple au moyen d'un k-means sur les coordonnées. A cela s'ajoute la difficulté d'interpréter les classes dans le nouvel espace factoriel. Avec notre méthode, les hiérarchies des classes permettent de choisir manuellement le niveau de détail souhaité en fonction du pourcentage d'information expliquée et nous pouvons alors facilement distinguer les classes les plus importantes en fonction de leur apport en information mutuelle.

6 Conclusion

Dans cet article, nous avons proposé une méthodologie d'utilisation du coclustering pour l'analyse exploratoire dans le cas de données mixtes. Un nombre de parties par variable étant fixé par l'analyste, les variables numériques sont discrétisées en intervalles de fréquence égale et les valeurs les plus fréquentes de variables catégorielles sont conservées. Un coclustering entre les individus et les variables ainsi binarisées est alors effectué, en laissant l'algorithme inférer automatiquement la taille de la matrice résumant le jeu de données. Nous avons montré sur un jeu de données de petite taille que l'analyse exploratoire fait apparaître un très bon accord entre ACM et coclustering, en dépit des différences d'approche tant du point de vue des modèles que des méthodologies utilisées. Nous avons montré que l'analyse exploratoire restait praticable sur un jeu de données plus complexe et de plus grande taille, en permettant une interprétation aisée du jeu de données par une analyse conjointe des clusters d'individus et de parties de variables. Les résultats de ces expérimentations sont particulièrement prometteurs et permettent déjà une utilisation en pratique de cette méthodologie sur des cas réels d'analyse exploratoire.

Néanmoins, cette méthode reste limitée par le choix d'un paramètre utilisateur : le nombre de parties par variable utilisé pour la binarisation du jeu de données. De plus, la méthode de coclustering utilisée n'exploite pas l'origine des parties, en particulier la structure de corrélation intrinsèque entre les parties d'une même variable qui forment une partition. Dans des travaux futurs, nous viserons à remédier à ces limites en définissant des modèles de coclustering intégrant le paramètre de granularité des binarisations et la connaissance des groupes de

colonnes formant des partition de variables. En définissant alors un critère d'évaluation spécialisé d'un tel coclustering ainsi que des algorithmes dédiés, nous espérons automatiser le choix de la granularité et améliorer la qualité des résultats.

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. (2008). Khiops : outil de préparation et modélisation des données pour la fouille des grandes bases de données. In *Extraction et gestion des connaissances*, pp. 229–230.

Boullé, M. (2011). Data grid models for preparation and modeling in supervised learning. In I. Guyon, G. Cawley, G. Dror, et A. Saffari (Eds.), *Hands-On Pattern Recognition : Challenges in Machine Learning*, pp. 99–130. Microtome Publishing.

Brault, V. et A. Lomet (2015). Revue des méthodes pour la classification jointe des lignes et des colonnes d'un tableau. *Journal de la Société Française de Statistique 156*(3), 27–51.

Dhillon, I. S., S. Mallela, et D. S. Modha (2003). Information-theoretic co-clustering. In *Proceedings of the ninth international conference on Knowledge discovery and data mining*, pp. 89–98. ACM Press.

Good, I. J. (1965). Categorization of classification. In *Mathematics and Computer Science in Biology and Medicine*, pp. 115–125. Her Majesty's Stationery Office, London.

Govaert, G. (1983). *Classification croisée*. Thèse d'état, Université Paris 6, France.

Govaert, G. et M. Nadif (2008). Block clustering with Bernoulli mixture models : Comparison of different approaches. *Computational Statistics and Data Analysis 52*(6), 3233–3245.

Govaert, G. et M. Nadif (2013). *Co-Clustering*. ISTE Ltd and John Wiley & Sons Inc.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York, NY, USA : John Wiley & Sons, Inc.

Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.

Summary

The cross-classification method is an unsupervised analysis technique that extracts the existing underlying structure between individuals and the variables in a data table as homogeneous blocks. This technique is limited to variables of the same type, either numerical or categorical, and we propose to extend it by proposing a two-step methodology. In the first step, all the variables are binarized according to a number of bins chosen by the analyst, by discretization in equal frequency in the numerical case, or keeping the most frequent values in the categorical case. The second step applies a coclustering method between the individuals and the binary variables, leading to groups of individual and groups of variable parts. We apply this methodology on several data sets and compare with the results of a multiple correspondence analysis MCA applied to the same data.