Pondération de blocs de variables en bi-partitionnement topologique

Amine Chaibi, Hanane Azzag, Mustapha Lebbah

{prenom.nom}@lipn.univ-paris13.fr
Université Paris 13, Sorbonne Paris Cité - CNRS
LIPN-UMR 7030
99, av. J-B Clément - F-93430 Villetaneuse

Résumé. Dans cet article, nous proposons une nouvelle approche permettant à la fois le bi-partitionnement topologique (bi-clustering) et la pondération de blocs variables. Le modèle que nous proposons FBR-BiTM (Feature Block Relevance using BiTM) permet de découvrir un espace topologique d'un ensemble d'observations et de variables en associant un nouveau score de pondération à chaque sous ensemble de variables. L'estimation des coefficients de pondération est réalisée dans le même processus d'apprentissage que le bi-partitionnement. Ces pondérations sont locales et associées à chaque prototype. Elles reflètent l'importance locale de chaque bloc de variables pour le bi-partitionnement. L'évaluation montre que l'approche proposée, comparée à d'autres méthodes de bi-partitionnement, obtient des résultats performants.
Mots clés : bi-partitionnement, pondérations de blocs variables, cartes topologiques.

1 Introduction

Les approches de bi-partitionnement sont devenues un sujet d'intérêt majeur en raison de leurs nombreuses applications dans le domaine de la fouille des données. Une méthode de bi-partitionnement, aussi appelée bi-clustering, co-clustering ou classification croisée, est une méthode d'analyse qui vise à regrouper des données en fonction de leur similarité. La stratégie classique des méthodes de bi-partitionnement cherche à trouver des sous-matrices ou des blocs, qui représentent des sous-groupes de lignes et des sous-groupes de colonnes d'une matrice de données.

Un des objectifs d'une méthode de bi-partitionnement est la recherche d'un couple de partitions, l'une sur les observations (les lignes d'une matrice de données), l'autre sur les variables (colonnes d'une matrice de données), tel que la "perte d'information" due au regroupement soit minimale (Charrad et al., 2008) ; c'est-à-dire de sorte que la différence entre l'information apportée par la matrice de données initiale et celle apportée par le regroupement obtenu soit minimale. Depuis le premier algorithme de bi-partitionnement, appelé Block Clustering proposé par Hartigan (1972), de nombreuses techniques ont été proposées telles que l'énumération exhaustive (Tanay et al., 2002), l'analyse spectrale (Greene et Cunningham, 2010), les réseaux

bayésiens (Shan et al., 2010) et d'autres (Angiulli et al., 2006). La pondération de variables est un processus couramment utilisé dans le domaine de l'apprentissage non supervisé, dont le but est de pondérer (ou sélectionner) des variables à partir d'une base de données en appliquant un algorithme d'apprentissage. La pondération de variables est difficile car, contrairement à l'apprentissage supervisé, les données ne sont pas étiquetées (Guyon et Elisseeff, 2003; Tsai et al., 2012). La taille des bases de données pose un défi sans précédent pour la fouille de données massives. Afin de remédier au problème de la grande dimension des variables, nous proposons dans un cadre de bi-partitionnement, une pondération des sous-ensembles de variables au lieu de pondérer les variables séparément.

2 Etat de l'art

Dans le domaine de la classification, bien que la plupart des méthodes utilisées cherchent à construire des partitions soit sur l'ensemble des observations soit sur celui des variables séparément. Il existe d'autres méthodes de bi-partitionnement qui considèrent simultanément les deux ensembles (Hartigan, 1972; Govaert, 1983; Nadif et Govaert, 2010; Ayadi et al., 2012). Les méthodes de bi-partitionnement utilisant les cartes auto-organisatrices (SOM) (Kohonen et al. (2001)) ont été définies par plusieurs auteurs (Busygin et al., 2002; Cottrell et al., 2004; Benabdeslem et Allab, 2012). Ce type de méthodes rentrent dans la catégorie des approches basées sur le partitionnement car souvent, elles utilisent des algorithmes de classification simple appliqués séparément sur les lignes et les colonnes d'une matrice des données. (Govaert, 1983) a défini un algorithme de bi-partitionnement nommé "Croeuc" pour les données quantitatives et qui consiste à déterminer une série de couples de partitions minimisant une fonction de coût sur la matrice des données en appliquant l'algorithme des nuées dynamiques. (Long et al., 2005) ont proposé une approche de décomposition matricielle "NBVD" (Non-negative Block Value Decomposition) pour le bi-clustering. Cette approche permet de décomposer une matrice de données en trois composantes en procédant par un algorithme itératif appliqué sur des données non négatives. Dans la même catégorie de méthodes, (Labiod et Nadif, 2011) ont proposé une approche de factorisation matricielle appelée "CUNMTF" (Co-clustering under Nonnegative Matrix Tri-Factorization). L'idée principale de cette approche est que la structure du bloc latent dans une matrice de données rectangulaire non négative est factorisée en deux facteurs plutôt que trois : la matrice des coefficients des lignes et la matrice des coefficients des colonnes qui indiquent respectivement le degré d'appartenance d'une ligne et d'une colonne à un cluster. On retrouve dans la littératures plusieurs approches de bi-partitionnement, qui utilisent des algorithmes hiérarchiques (Caldas et Kaski, 2011; Mao et al., 2005; Getz et al., 2000a). L'approche la plus utilisée dans cette famille de modèle est CTWC (Coupled two-way clustering) (Getz et al., 2000a). CTWC consiste à appliquer un algorithme de classification hiérarchique, le SPC "Super Paramagnetic Clustering (SPC)" (Getz et al., 2000b) sur les colonnes en utilisant toutes les lignes puis sur les lignes en utilisant toutes les colonnes.

Dans ce papier, une approche de pondération de blocs de variables en utilisant un modèle de bi-partitionnement, basée sur les cartes topologiques est proposée. Plusieurs méthodes de pondération de variables sont recensées dans la littérature scientifique. Nous trouvons des approches de pondération locale basées sur l'apprentissage non supervisé (Blansché et al., 2006; Frigui et Nasraoui, 2004) ainsi que sur les k-means (Huang et al., 2005). Il existe aussi des méthodes de sélection locale de caractéristiques basées sur l'apprentissage non supervisé (Basak

et al., 1998; Liu et al., 2009; Grozavu et al., 2009; Chen et al., 2012). (Ouattara et al., 2013) proposent une extension des cartes topologiques pour le traitement des données multiblocs.

Dans ce papier, nous proposons une approche qui permet d'aborder le problème de la pondération de "blocs de variables" dans un cadre de bi-clustering topologique. Pour cela, nous nous basons sur l'algorithme BiTM (Biclustering using Topological Maps) (Chaibi et al., 2013). Notre modèle attribue à chaque bloc de variables un nouveau score de pondération constituant un vecteur des pondérations locale nommé FBR (Feature Block Relevance). La principale différence entre notre approche nommée FBR_BiTM et les méthodes existantes est que la pondération n'est pas associée à une seule variable, mais à un bloc de variables.

3 Bi-partitionnement et pondération des blocs de variables

Le modèle FBR_BiTM est constitué d'un ensemble de cellules discrètes $\mathcal{C}$ de taille K appelées "carte". Pour chaque paire de cellules (c, r) de la carte, la distance $\delta(c, r)$ est définie par le plus court chemin reliant les cellules r et c sur la grille. Soit $\Re^d$ l'espace euclidien des données et $\mathcal{A}$ la matrice des données où chaque observation $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^j, \ldots, x_i^d)$ est un vecteur dans $\Re^d$. L'objectif de FBR_BiTM est de fournir des bi-clusters organisés dans une carte topologique et un vecteur des pondérations locales de chaque bloc de variables. Dans FBR_BiTM, chaque cellule c de $\mathcal{C}$ est associée à un prototype sous la forme d'un vecteur : $\mathbf{g}_k = (g_k^1, g_k^2, \ldots, g_k^l, \ldots, g_k^L)$ et un vecteur des pondérations $\mathbf{f_k} = (f_k^1, f_k^2, \ldots, f_k^l, \ldots, f_k^L)$ de dimension $L < d$ où g_k^l est la valeur représentante du bloc B_k^l et f_k^l la valeur de pondération locale. L'ensemble des lignes (observations) $I = \{1, \ldots, N\}$ de la matrice des données $\mathcal{A}$ est partitionné en K groupes $\{P_1, P_2, \ldots, P_k, \ldots, P_K\}$. De même, l'ensemble des colonnes (variables) $J = \{1, \ldots, d\}$ est partitionné en L groupes $\{Q_1, Q_2, \ldots, Q_l, \ldots, Q_L\}$. Nous définissons deux matrices binaires $Z = (z_i^k)$ et $W = (w_j^l)$ pour sauvegarder les informations associées respectivement aux observations et aux variables.

$$z_i^k = \begin{cases} 1 & \text{si } \mathbf{x}_i \in P_k, k = \phi_z(\mathbf{x}_i) \\ 0 & \text{sinon} \end{cases}$$

$$w_j^l = \begin{cases} 1 & \text{si } \mathbf{x}^j \in Q_l, l = \phi_w(\mathbf{x}^j) \\ 0 & \text{sinon} \end{cases}$$

Où ϕ est la fonction d'affectation. Avec z_i^k et w_j^l, nous pouvons déterminer des blocs de variables $B_k^l = \{x_i^j | z_i^k \times w_j^l = 1\}$. Le modèle que nous proposons dans ce papier se base sur la formulation du modèle BiTM (Chaibi et al., 2013) en introduisant le nouveau paramètre f_r^l qui sera estimé au cours de l'apprentissage. Ainsi nous proposons de minimiser la nouvelle fonction de coût suivante :

$$\mathcal{J}(\phi_w, \phi_z, G, F) = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{N}\sum_{j=1}^{d}\sum_{r=1}^{K} \mathcal{K}^T(\delta(r,k)) \times w_j^l \times z_i^k \times (f_r^l \times x_i^j - g_r^l)^2$$

Cette fonction de coût peut être réécrite de la manière suivante :

$$\mathcal{J}(\phi_w, \phi_z, G, F) = \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{\mathbf{x}_i \in P_k}\sum_{\mathbf{x}^j \in Q_l}\sum_{r=1}^{K} \mathcal{K}^T(\delta(r,k)) \times (f_r^l \times x_i^j - g_r^l)^2$$

Où :

$G = \{\mathbf{g_1}, \ldots, \mathbf{g_k}\}$ désigne l'ensemble des vecteurs prototypes.
$F = \{\mathbf{f_1}, \ldots, \mathbf{f_k}\}$ représente l'ensemble des vecteurs de pondération.
ϕ_z est la fonction d'affectation des lignes.
ϕ_w est la fonction d'affectation des colonnes.
$\mathcal{K}^T(\delta(r,k))$ est la fonction de voisinage. En pratique, nous utilisons la fonction de voisinage suivante : $\mathcal{K}^T(\delta(c,r)) = \exp\left(\frac{-\delta(c,r)}{T}\right)$ où T représente le paramètre contrôlant le rayon du voisinage.

La minimisation de $\mathcal{J}(\phi_w, \phi_z, G, F)$ se fait d'une manière itérative avec la version nuées dynamiques par l'exécution de 4 étapes jusqu'à un nombre d'itérations prédéfini (algorithme 1). De la même manière que les cartes topologiques, on fait décroître le rayon d'apprentissage pour constituer deux phase : une phase d'auto-organisation associée aux grandes valeurs et une phase de quantification associée aux petites valeurs.

4 Expérimentations

Nous avons testé l'algorithme FBR_BiTM avec des jeux de données du répertoire UCI (Frank et Asuncion (2010)). Nous avons aussi utilisé dans ces expérimentations des bases de données binaires synthétiques [1]. La particularité de ces bases de données est que les observations et les variables sont étiquetées. Nous allons particulièrement utiliser ces bases de données afin d'évaluer le clustering des variables. Les tableaux 1 et 2 indiquent les paramètres de chaque jeu de données (nombre d'observations, nombre de variables, la taille de la carte utilisée pour l'apprentissage et le nombre de classes réelles).

Bases de données	# Observations	# Variables	Taille carte	# Classes
isolet5	1559	617	12×12	26
Breast	699	10	7×7	2
Sonar Mines	208	60	6×6	2
Lung Cancer	32	56	4×4	2
Spectf 1	349	44	4×4	2
Cancer Wpbc Ret	198	33	6×6	2
Horse Colic	300	27	5×5	2
Heart	270	13	5×5	2
glass	214	9	5×5	7

TAB. 1 – *Description des jeux de données du site UCI.*

1. Ces bases de données sont fournis par Dr. Lazhar Labiod https ://sites.google.com/site/lazharlabiod/

Algorithme 1 Algorithme FBR_BiTM

ENTRÈES : Les données $\mathcal{A} = \{x_i^j\}_{i=1...N, j=1...d}$. Les matrices d'affectation Z, W. Les prototypes G de la carte initialisées. Les vecteurs de pondération F initialisées. t_{max} : le nombre maximum d'itérations.

SORTIES : Les matrices d'affectation Z, W. Les prototypes G mis à jour. Les vecteurs de pondération F mis à jour.

Phase itérative

1- Affectation des observations : chaque observation $\mathbf{x}_i$ est affectée au prototype $\mathbf{g}_k$ le plus proche en utilisant la fonction d'affectation :

$$\phi_z(\mathbf{x}_i) = \arg\min_c \sum_{j=1}^{d} \sum_{l=1}^{m} \sum_{r=1}^{K} w_j^l \times \mathcal{K}^T(\delta(r,c)) \times (f_r^l \times x_i^j - g_r^l)^2$$

2- Affectation des variables : chaque variable $\mathbf{x}^j$ est affectée au prototype le plus proche en utilisant la fonction d'affectation :

$$\phi_w(\mathbf{x}^j) = \arg\min_l \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{r=1}^{K} z_i^k \times \mathcal{K}^T(\delta(r,k)) \times (f_r^l \times x_i^j - g_r^l)^2$$

3- Mise à jour des prototypes : les composantes g_r^l des prototypes sont mis à jour suivant la formule ci-dessous :

$$g_r^l = \frac{\sum_{i=1}^{N} \sum_{j=1}^{d} \mathcal{K}^T(\delta(r, \phi_z(\mathbf{x}_i))) \times w_j^l \times x_i^j \times f_r^l}{\sum_{i=1}^{N} \sum_{j=1}^{d} \mathcal{K}^T(\delta(r, \phi_z(\mathbf{x}_i))) \times w_j^l}$$

4- Mise à jour des pondérations : les composantes f_r^l des vecteurs des pondérations sont mis à jour suivant la formule ci-dessous :

$$f_r^l = \frac{\sum_{i=1}^{N} \sum_{j=1}^{d} \mathcal{K}^T(\delta(r, \phi_z(\mathbf{x}_i))) \times w_j^l \times x_i^j \times g_r^l}{\sum_{i=1}^{N} \sum_{j=1}^{d} \mathcal{K}^T(\delta(r, \phi_z(\mathbf{x}_i))) \times w_j^l \times (x_i^j)^2}$$

RÉPÉTER les phases 1, 2, 3 et 4 jusqu'à $t = t_{max}$.

Bases de données	# Observations	# Variables	Taille carte	# Classes (obs/var)
Simulé 1	2000	5000	4×4	3
Simulé 2	2000	5000	6×6	3
Simulé 3	2000	5000	8×8	3
Simulé 4	2000	5000	10×10	3

TAB. 2 – *Description des jeux de données simulées.*

Protocole de validation

Afin de comparer FBR-BiTM avec les approches de bi-partitionnement, nous avons sélectionné les approches suivantes : BiTM (Chaibi et al. (2013)), CTWC (Getz et al. (2000a)), NBVD (Long et al. (2005)). Les résultats expérimentaux sont présentés dans les tableaux 3 et 4. Nous avons choisi la taille des cartes FBR-BiTM et BiTM selon l'heuristique de Kohonen. Le nombre de clusters des variables (colonnes de la matrice $\mathcal{A}$) est exactement le même pour l'ensemble des approches FBR-BiTM, BiTM, CTWC, NBVD. Cependant, pour le nombre de clusters des observations (lignes de la matrice $\mathcal{A}$), nous avons pris la même taille des cartes FBR-BiTM et BiTM, par contre pour les approches CTWC et NBVD, nous avons pris une taille proportionnelle au nombre de cellules non vides dans FBR-BiTM et BiTM. Par exemple, dans le cas de la base simulée 1, la taille de la carte FBR-BiTM est identique à la taille de la carte BiTM qui est égale $4 \times 4 = 16$. Le nombre de cellules vides des cartes FBR-BiTM et BiTM est égale à 4. Ainsi, la taille de la partition des lignes que nous avons choisi pour les approches CTWC et NBVD est égale à 16-4 = 12. L'initialisation des partitions des lignes et des colonnes est effectuée d'une manière aléatoire pour l'ensemble des approches FBR-BiTM, BiTM, CTWC et NBVD. Nous avons normalisé l'ensemble des jeux de données entre 0 et 1 Nous avons calculé deux indices de performances (pureté et rand) sur l'ensemble des résultats obtenus avec les approches FBR-BiTM, BiTM, CTWC et NBVD. Nous avons sélectionné la meilleure performance obtenue dans les 10 expérimentations réalisées.

4.1 Comparaison des performances de FBR-BiTM avec les approches de bi-partitionnement

Le tableau 3 résume les résultats expérimentaux de l'indice de pureté. Nous remarquons que FBR-BiTM fournit des résultats équivalents à ceux de l'approche BiTM. Dans la plupart des cas, nous constatons des résultats comparable et souvent meilleur avec notre approche BiTM ou FBR-BiTM. Nous observons aussi la difficulté d'obtenir de grandes valeurs de l'indice de pureté pour la base isolet5. Le tableau 4 présente l'indice de rand. Nous observons que FBR-BiTM fournit un indice de rand meilleur que celui des autres approches dans 4 bases sur 9. Les résultats de FBR-BiTM restent compétitifs et équivalents aux résultats obtenus avec les autres approches. Nous constatons après cette étude comparative, que notre approche FBR-BiTM est une méthode qui ne perturbe pas le bi-partitionnement topologiques BiTM.

Base de données	FBR-BiTM	BiTM	CTWC	NBVD
isolet5	**0.441**	0.316	0.103	0.073
Breast	0.968	**0.978**	0.655	0.834
Sonar Mines	**0.770**	0.769	0.548	0.644
Lung Cancer	0.843	**1**	0.718	0.875
Spectf 1	**0.767**	0.759	0.727	0.727
Cancer Wpbc Ret	0.772	**0.787**	0.762	0.762
Horse Colic	**0.723**	0.719	0.67	0.67
Heart	0.814	**0.883**	0.555	0.674
glass	0.439	**0.618**	0.523	0.462

TAB. 3 – *Bi-partitionnement : comparaison en utilisant l'indice de pureté obtenu avec FBR-BiTM, BiTM, CTWC et NBVD*

Base de données	FBR-BiTM	BiTM	CTWC	NBVD
isolet5	0.858	**0.926**	0.91	0.502
Breast	**0.790**	0.687	0.505	0.659
Sonar Mines	0.494	0.508	0.502	**0.514**
Lung Cancer	**0.687**	0.459	0.556	0.556
Spectf 1	0.416	0.418	**0.513**	0.42
Cancer Wpbc Ret	0.435	0.435	**0.524**	0.414
Horse Colic	**0.474**	**0.472**	0.463	0.46
Heart	**0.615**	0.56	0.498	0.513
glass	0.607	0.653	0.69	**0.693**

TAB. 4 – *Bi-partitionnement : comparaison en utilisant l'indice de Rand obtenu avec FBR-BiTM, BiTM, CTWC et NBVD*

Cas particulier : application aux bases de données simulées binaires

Dans les bases de données réelles, il est très difficile d'obtenir les étiquettes des classes des variables. Afin de valider le clustering des variables de notre modèle FBR-BiTM, nous avons utilisé des bases de données simulées étiquetées en lignes (observations) et en colonnes (variables) décrites dans le tableau 2. Les tableaux 5 et 6 montrent les résultats obtenus avec les indices de pureté et de rand pour le partitionnement des observations et des variables.

Nous constatons à travers les 2 indices de performance que notre approche est meilleure ou équivalente à BiTM dans le cas du clustering des observations dans la plupart des bases de données. Cependant, nous remarquons une légère baisse des performances de FBR-BiTM au niveau du clustering des variables.

4.2 Résultats visuels

Dans cette partie, nous montrons l'apport visuel de FBR-BiTM. Notre approche FBR-BiTM se base sur les visualisations intuitives des cartes auto-organisatrices. Les figures 1 et 2 représentent différentes visualisations obtenues sur les bases de données simulée1 et Lung Cancer. Les figures 1(a) et 2(a) sont dédiées à la visualisation de la base de données organisées

Bases	FBR-BiTM		BiTM		CTWC		NBVD	
	Obser	Var	Obser	Var	Obser	Var	Obser	Var
Simulé 1	0.983	**0.699**	**0.991**	**0.704**	0.901	0.512	0.828	0653
Simulé 2	**0.901**	0.762	0.881	**0.794**	**0.902**	0.612	0.712	0.783
Simulé 3	0.983	**0.799**	**0.999**	**0.793**	0.910	0.692	0.623	0.483
Simulé 4	**0.901**	0.703	0.881	**0.781**	**0.891**	0.732	0.391	0.802

TAB. 5 – *Indice de pureté sur les bases simulées binaires des approches FBR-BiTM, BiTM, CTWC, NBVD.*

Bases	FBR-BiTM		BiTM		CTWC		NBVD	
	Obser	Var	Obser	Var	Obser	Var	Obser	Var
Simulé 1	**0.928**	0.403	**0.930**	0.441	0.910	0.412	0.492	**0.510**
Simulé 2	0.803	0.391	**0.889**	0.370	0.718	**0.409**	0.831	0.352
Simulé 3	**0.882**	0.536	**0.875**	0.363	0.682	0.282	0.721	0.401
Simulé 4	0.853	0.293	**0.889**	0.370	0.812	**0.512**	0.691	0.290

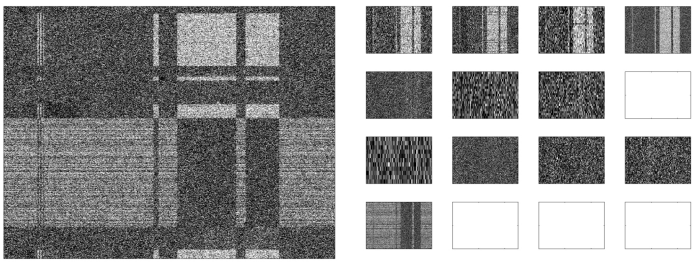
TAB. 6 – *Indice de rand sur les bases simulées binaires des approches FBR-BiTM, BiTM, CTWC, NBVD et CUNMTF.*

en fonction des groupes de lignes et de colonnes. Cette organisation est très claire dans le cas des bases binaires (voir la figure 1(a)).

Ces figures peuvent être obtenues par toutes les méthodes de bi-partitionnement. Cependant, en utilisant cette visualisation, il est difficile d'analyser les blocs ou les bi-clusters obtenus. Afin de faciliter cette tâche, nous proposons de visualiser les bi-clusters en utilisant l'organisation topologique du modèle FBR-BiTM. Ainsi, chaque cellule de la carte est associée au cluster des observations et des variables. Cette organisation est illustrée par les figures 1(b) et 2(c) en organisant les cellules selon l'ordre des blocs de variables obtenus. Dans le cas de la base Lung Cancer par exemple, la figure 2(b) représente la carte topologique associée au modèle FBR-BiTM. Cette figure représente la topologie des groupes obtenus en appliquant l'algorithme FBR-BiTM. Nous remarquons une répartition des données au niveau de chaque cellule. Plus la couleur est rouge, plus les variables ont de fortes valeurs. Nous avons organisé la carte selon les blocs de variables obtenus. Le résultat est illustré dans la figure 2(c). Dans la première cellule par exemple, nous remarquons que les variables ont changé de disposition de manière à créer une organisation au niveau de la cellule. Nous constatons clairement dans cette première cellule (en haut à gauche de la carte) que les blocs de variables se comportent différemment à l'intérieur de cette cellule. Ce comportement est illustré par une couleur. Plus la couleur est rouge, plus le bloc de variable tend vers de fortes valeurs. Dans ce cas, nous remarquons que les premiers blocs de variables (totalement à gauche de la cellule) ont une couleur plutôt rouge. Par contre, le second bloc (au milieu de la cellule) est moins "important" car il est constitué de variables d'une couleur bleu. Enfin, le troisième bloc (totalement à droite) est constitué des variables moyennement importantes (couleur verte). Nous nous sommes focalisés dans cette analyse sur la base Lung Cancer. En fait, cette analyse peut être également réalisée sur les autres bases de données.

Distribution des blocs de variables : la figure 2(d) représente les blocs de variables sur

la carte. Ces visualisations sont intéressantes car elles permettent de représenter un résumé exhaustif de l'ensemble des variables. En effet, au lieu de visualiser toutes les variables de la base Lung Cancer (56 variables) par exemple, on ne visualise que 4 blocs de variables. A partir de cette visualisation aussi, nous disposons d'une information sur la variations des blocs de variables sur toute la carte. Il est clair que les blocs 1 et 3 de la base Lung Cancer varient de la même manière. Finalement FBR-BiTM a l'avantage de proposer une visualisation de la base de données et des bi-clusters. Ce résultat permet aux utilisateurs/experts une meilleure compréhension de la cohérence des données.



(a) La base de données organisées en fonction de l'ordre des observations et des variables de la classification croisée.

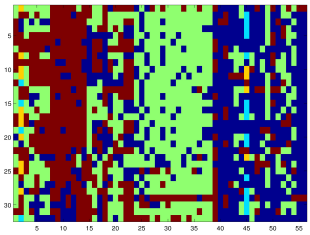
(b) Carte FBR-BiTM organisée selon les blocs de variables.

FIG. 1 – *Visualisation de la base de données simulées 1 (binaire) en utilisant FBR-BiTM. Chaque cellule dans la figure 1(b) indique une cellule de la carte.*

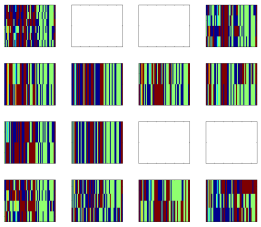
5 Conclusion et perpectives

Nous avons présenté dans ce papier une nouvelle approche de pondération de "blocs de variables" en utilisant un modèle de bi-partitionnement topologique. La principale nouveauté du modèle FBR-BiTM est l'utilisation d'un modèle topologique pour organiser la matrice de données, en blocs homogènes en prenant en compte simultanément les lignes et les colonnes, et l'apprentissage d'un nouveau critère d'importance de chaque blocs de variables. La série d'expériences que nous avons réalisé, a permis de valider notre méthode et d'analyser ses performances. Notre algorithme offre de nouvelles visualisations permettant de mieux comprendre la structure de données. Il est clair que nous allons améliorer le modèle FBR-BiTM afin d'obtenir de meilleurs performances du clustering des observations et des variables. Nous souhaitons aussi conforter les expérimentations avec d'autres bases de données et d'autres algorithmes.

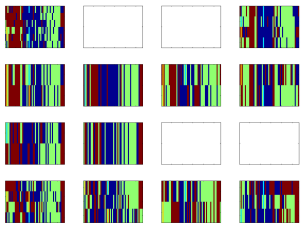
Remerciement : Ce travail est réalisé dans le cadre du projet Square Predict (investissement d'avenir), financé par le fonds national pour la société numérique (FSN).



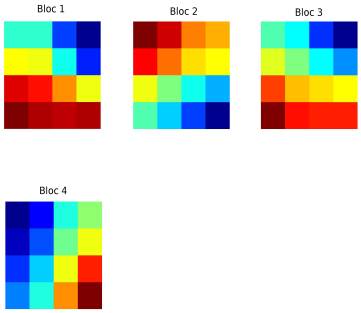
(a) La base de données organisées en fonction de l'ordre des observations et des variables de la classification croisée.



(b) La carte FBR-BiTM. Représentation topologique des groupes de la carte FBR-BiTM.



(c) Carte FBR-BiTM organisée selon les blocs de variables



(d) Représentation des blocs de variables

FIG. 2 – *Visualisation de la base de données Lung Cancer en utilisant FBR-BiTM. Chaque cellule dans la figure 2(c) indique une cellule de la carte.*

Références

Angiulli, F., E. Cesario, et C. Pizzuti (2006). A greedy search approach to co-clustering sparse binary matrices. In *ICTAI*, pp. 363–370. IEEE Computer Society.

Ayadi, W., M. Elloumi, et J.-K. Hao (2012). Pattern-driven neighborhood search for biclustering of microarray data. *BMC Bioinformatics 13*(S-7), S11.

Basak, J., R. K. De, et S. K. Pal (1998). Unsupervised feature selection using a neuro-fuzzy approach. *Pattern Recognition Letters 19*(11), 997–1006.

Benabdeslem, K. et K. Allab (2012). Bi-clustering continuous data with self-organizing map. Neural Computing and Applications.

Blansché, A., P. Gançarski, et J. J. Korczak (2006). Maclaw : A modular approach for clustering with local attribute weighting. *Pattern Recognition Letters 27*(11), 1299–1306.

Busygin, S., G. Jacobsen, E. Kremer, et C. Ag (2002). Double conjugated clustering applied to leukemia microarray data. In *In 2nd SIAM ICDM, Workshop on clustering high dimensional data*.

Caldas, J. et S. Kaski (2011). Hierarchical generative biclustering for microrna expression analysis. *Journal of Computational Biology 18*(3), 251–261.

Chaibi, A., M. Lebbah, et H. Azzag (2013). Nouvelle approche de bi-partitionnement topologique. In *EGC*, Volume RNTI-E-24, pp. 37–48. Hermann-Éditions.

Charrad, M., G. Saporta, Y. Lechevallier, et M. Ben Ahmed (2008). Le bi-partitionnement : Etat de l'art sur les approches et les algorithmes. In *Ecol'IA 2008*.

Chen, X., Y. Ye, X. Xu, et J. Z. Huang (2012). A feature group weighting method for subspace clustering of high-dimensional data. *Pattern Recognition 45*(1), 434 – 446.

Cottrell, M., S. Ibbou, et P. Letrémy (2004). Som-based algorithms for qualitative variables. *Neural Netw. 17*(8-9), 1149–1167.

Frank, A. et A. Asuncion (2010). Uci machine learning repository. *Technical report, School of Information and Computer Sciences, available at :http ://archive.ics.uci.edu/ml*.

Frigui, H. et O. Nasraoui (2004). Unsupervised learning of prototypes and attribute weights. *Pattern Recognition 37*(3), 567–581.

Getz, G., E. Levine, et E. Domany (2000a). Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA 97*, 12079–12084.

Getz, G., E. Levine, E. Domany, et M. Q. Zhang (2000b). Super paramagnetic clustering of yeast gene expression profiles.

Govaert, G. (1983). *Classification croisée*. Ph. D. thesis, Université Paris 6, France.

Greene, D. et P. Cunningham (2010). Spectral co-clustering for dynamic bipartite graphs. In *Workshop on dynamic networks and knowledge discovery at ecml'10, barcelona, spain*.

Grozavu, N., Y. Bennani, et M. Lebbah (2009). From variable weighting to cluster characterization in topographic unsupervised learning. In *International Joint Conference on Neural Networks, IJCNN 2009, Atlanta, Georgia, USA, 14-19 June 2009*, pp. 1005–1010.

Guyon, I. et A. Elisseeff (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, 1157–1182.

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

Huang, J. Z., M. K. Ng, H. Rong, et Z. Li (2005). Automated variable weighting in k-means type clustering. *IEEE Trans. Pattern Anal. Mach. Intell. 27*(5), 657–668.

Kohonen, T., M. R. Schroeder, et T. S. Huang (Eds.) (2001). *Self-Organizing Maps* (3rd ed.). Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Labiod, L. et M. Nadif (2011). Co-clustering under nonnegative matrix tri-factorization. In *Proceedings of the 18th international conference on Neural Information Processing - Volume Part II*, ICONIP'11, Berlin, Heidelberg, pp. 709–717. Springer-Verlag.

Liu, R., N. Yang, X. Ding, et L. Ma (2009). An unsupervised feature selection algorithm: laplacian score combined with distance-based entropy measure. In *Proceedings of the 3rd international conference on Intelligent information technology application*, IITA'09, Piscataway, NJ, USA, pp. 65–68. IEEE Press.

Long, B., Z. M. Zhang, et P. S. Yu (2005). Co-clustering by block value decomposition. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, New York, NY, USA, pp. 635–640. ACM.

Mao, D., Y. Luo, J. Zhang, et J. Zhu (2005). A new strategy of cooperativity of biclustering and hierarchical clustering: a case of analyzing yeast genomic microarray datasets. *Front Biosci 10*.

Nadif, M. et G. Govaert (2010). Model-based co-clustering for continuous data. In *ICMLA 2010 Proceedings ICMLA 2010, The Ninth International Conference on Machine Learning and Applications*, Washington États-Unis, pp. 1–6.

Ouattara, M., N. N. Keita, F. Badran, et C. Mandin (2013). Soft Subpace clustering pour données multiblocs basée sur les cartes topologiques auto-organisées SOM : 2S-SOM. In *SFDS 2013*, Toulouse,.

Shan, H., , et A. Banerjee (2010). Residual bayesian co-clustering for matrix approximation. In *SDM*, pp. 223–234.

Tanay, A., R. Sharan, et R. Shamir (2002). Discovering statistically significant biclusters in gene expression data. In *In Proceedings of ISMB 2002*, pp. 136–144.

Tsai, C.-A., C.-H. Huang, C.-W. Chang, et C.-H. Chen (2012). Recursive feature selection with significant variables of support vectors. *Comp. Math. Methods in Medicine 2012*.

Summary

The model that we propose named FBR-BITM (Feature Block Relevance using BiTM) allows to discover a topological space of observation and features by associating a new score of weighting for each feature subset. The weights is estimated in the same learning process of bi-clustering algorithm. These weights are local and associated with each prototype. They reflect the local relevance of each feature block for bi-clustering. The evaluation results show that the proposed methods compared to other bi-clustering approaches are effective.
Keywords : bi-clustering, feature block weighting, topological maps.