Nouvelle approche de bi-partitionnement topologique

Amine Chaibi[*,**], Mustapha Lebbah[*], Hanane Azzag[*]

[*] {prenom.nom}@lipn.univ-paris13.fr
[*]Université Paris 13, Sorbonne Paris Cité - CNRS
LIPN-UMR 7030
99, av. J-B Clément - F-93430 Villetaneuse
[**] Anticipeo
4 bis, impasse Courteline 94800 Villejuif, France

Résumé. Dans ce papier, nous proposons une nouvelle approche topologique de bi-partitionnement (bi-clustering) appelée BiTM en utilisant les cartes auto-organisatrices. L'idée principale de l'approche est d'utiliser une seule carte pour le partitionnement simultané des lignes (observations) et des colonnes (variables). Contrairement aux approches utilisant les cartes topologiques, notre modèle ne nécessite pas de pré-traitement de la base de données. Ainsi, une nouvelle fonction de coût est proposée. De plus, BiTM fournit une visualisation topologique des blocs ou bi-clusters facilement interprétable. Les résultats obtenus sont très encourageants et prometteurs pour continuer dans cette optique.

1 Introduction

Les approches de bi-partitionnement sont devenues un sujet d'intérêt en raison de ses nombreuses applications dans le domaine de l'exploration des données. Une méthode de bi-partitionnement, aussi appelée classification croisée, bi-clustering ou co-clustering, 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. 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),Charrad et al. (2008)). L'approche Block Clustering (Hartigan (1972)) permet de diviser la matrice des données en plusieurs sous-matrices correspondant à des blocs. Le principe de base de cette méthode est de faire des permutations des lignes et des colonnes afin de définir la structure de bloc. De plus, l'auteur Hartigan (1972) a proposé deux autres algorithmes de bi-partitionnement : le premier (One-Way Splitting) est principalement basé sur le partitionnement des observations en utilisant des fonctions ayant une variance intra-classe supérieure à un seuil donné afin de diviser la classe associée. Le second algorithme (Two-Way Splitting) procède par des divisions successives des lignes et des colonnes. Le même principe a été repris dans l'approche CTWC proposée par Getz et al.

(2000a). CTWC consiste à appliquer un algorithme de classification hiérarchique, le SPC (Super Paramagnetic Clustering) introduit par Getz et al. (2000b) sur les colonnes en utilisant toutes les lignes et vice versa.

Les algorithmes de k-means ont longuement été utilisés dans le bi-partitionnement. En effet, Govaert (1983) a défini trois algorithmes de bi-partitionnement : Croeuc, Crobin et Croki2. Ces algorithmes consistent à déterminer une série de couples de partitions minimisant une foction de coût sur la matrice des données en appliquant la méthode des nuées dynamiques alternativement sur les lignes et les colonnes. Croeuc, Crobin et Croki2 diffèrent par le type des données à traiter. En effet, Croeuc est destiné à des données quantitatives. Crobin est appliqué sur des données binaires. Croki2 est utilisé pour un tableau de contingence.

Récemment, de nouvelles approches de bi-partitionnement basées sur la décomposition matricielle sont proposées (Paatero et Tapper (1994), Long et al. (2005), Yoo et Choi (2010), Labiod et Nadif (2011), Shang et al. (2012)). Les auteurs de Long et al. (2005) ont proposé une approche nommée NBVD qui décompose une matrice des données en trois composantes en procédant par un algorithme itératif appliqué sur des données non négatives. L'approche nommée Coclustering Under Nonnegative Matrix Tri-Factorization (CUNMTF) introduite par Labiod et Nadif (2011) appartient à cette même famille. Les auteurs montrent que le double k-means est équivalent à un problème algébrique de NMF sous certaines contraintes appropriées.

Les méthodes de bi-partitionnement utilisant des cartes auto-organisatrices (SOM) (Kohonen et al. (2001)) ont été définis par plusieurs auteurs. Nous citons l'approche DCC (Double Conjugated Clustering) de Busygin et al. (2002) et KDISJ (Kohonen for Disjonctive Table) de Cottrell et al. (2004) ainsi qu'une autre variante récente introduite par Benabdeslem et Allab (2012). L'inconvénient de la méthode DCC est l'utilisation de deux cartes (une carte pour les observations et une carte pour les variables). Ces cartes sont construites indépendamment avec la même dimension. En ce qui concerne KDISJ, cette méthode est uniquement dédiée aux données catégorielles.

Dans ce papier, nous proposons une nouvelle approche (BiTM) de bi-partitionnement utilisant les cartes topologiques. BiTM ne nécessite aucune pré-organisation de la matrice des données en utilisant une seule carte qui représente simultanément la partition des observations et la partition des variables. Notre approche permet aussi de fournir de nouvelles visualisations. Le reste de cet article est organisé comme suit : dans la section 2, nous présentons le modèle et l'algorithme, la section 3 est dédiée à la méthodologie et les résultats expérimentaux. Enfin, nous concluons cet article par une conclusion et quelques perspectives.

2 Bi-partitionnement topologique : modèle BiTM

Le modèle BiTM est constitué d'un ensemble de cellules discrètes $\mathcal{C}$ de taille K appelées "carte". Cette carte a une topologie discrète définie comme un graphe non orienté, qui est généralement une grille à 2 dimensions. 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 D 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 BiTM est de fournir des bi-clusters organisés dans une carte topologique. Pour cela, l'ensemble des lignes (observations) $I = \{1, \ldots, N\}$ de la matrice des données D 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 $Z = (z_{ik})$ et $W = (w_{jl})$ pour sauvegarder les informations associées respectivement aux observations et aux variables.

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

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

Où ϕ est la fonction d'affectation. Avec z_{ik} et w_{jl}, nous pouvons déterminer des blocs de données $B_k^l = \{x_{ij} | z_{ik} \times w_{jl} = 1\}$. Dans BiTM, chaque cellule c de $\mathcal{C}$ est associée à un prototype sous la forme d'un vecteur : $\mathbf{g}_k = (g_k^1, g_k^2, \ldots, g_k^l, \ldots, g_k^L)$ de dimension $L < d$ où g_k^l est le prototype du bloc B_k^l. Nous proposons de minimiser la nouvelle fonction de coût suivante :

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

$G = \{\mathbf{g}_1, \ldots, \mathbf{g}_K\}$ désigne l'ensemble des prototypes.
ϕ_z la fonction d'affectation des lignes.
ϕ_w la fonction d'affectation des colonnes.
$\mathcal{K}^T(\delta(r,k))$ la fonction de voisinage.
T représente la fonction contrôlant le rayon du voisinage.
De même que pour les cartes auto-organisatrices, nous utilisons la fonction $\mathcal{K}^T(\delta(c,r)) = \exp(\frac{-\delta(c,r)}{T})$ pour définir le voisinage.

La minimisation de $\tilde{R}(\phi_w, \phi_z, G)$ est obtenue par l'exécution itérative de 4 étapes jusqu'à un nombre d'itérations prédéfini (algorithme 1).

2.1 L'ordre topologique dans le modèle BiTM

La décomposition de la fonction de coût $\tilde{R}$ qui dépend de la valeur de T, peut être réécrite de la manière suivante :

$$\begin{aligned} \tilde{R}(\phi_w, \phi_z, G) &= \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{\mathbf{x}_i \in P_k} \sum_{\mathbf{x}^j \in Q_l} \sum_{r=1, r \neq k}^{K} \mathcal{K}^T(\delta(r,k)) \parallel x_i^j - g_r^l \parallel^2 \\ &+ \mathcal{K}^T(0) \sum_{r=1}^{K} \sum_{l=1}^{L} \sum_{\mathbf{x}_i \in P_k} \sum_{\mathbf{x}^j \in Q_l} \parallel x_i^j - g_r^l \parallel^2 \end{aligned}$$

La fonction de coût $\tilde{R}$ est décomposée en deux termes. Afin de maintenir l'ordre topologique entre les blocs, la minimisation du premier terme entraîne le bloc qui correspond à deux cellules voisines. En effet, si les cellules c et r sont voisines dans la carte, la valeur de $\delta(r,k)$ est faible et dans ce cas, la valeur de $\mathcal{K}^T(\delta(r,k))$ est élevée. La minimisation du second terme correspond à la minimisation de l'inertie des données locales affectées à un bloc $B_r^j, j = 1 \ldots L$. Pour différentes valeurs de T, chaque terme de la fonction de coût a une importance relative dans le processus de minimisation. On peut, donc, définir deux étapes pour l'exploitation de l'algorithme :

Algorithme 1

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

SORTIES :
- Les matrices d'affectation Z, W.
- Les prototypes G 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 :

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

2- Mise à jour des prototypes : les vecteurs des prototypes sont mis à jour en fonction des affectations des observations :

$$g_k^l = \frac{\sum_{i=1}^{N} \sum_{j=1}^{d} \mathcal{K}^T(\delta(k, \phi_z(\mathbf{x}_i)) \times w_{jl} \times x_i^j}{\sum_{i=1}^{N} \sum_{j=1}^{d} \mathcal{K}^T(\delta(k, \phi_z(\mathbf{x}_i))) \times w_{jl}}$$

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

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

4- Mise à jour des prototypes : les vecteurs des prototypes sont mis à jour en fonction des affectations des variables :

$$g_k^l = \frac{\sum_{i=1}^{N} \sum_{j=1}^{d} \mathcal{K}^T(\delta(k, \phi_z(\mathbf{x}_i))) \times w_{jl} \times x_i^j}{\sum_{i=1}^{N} \sum_{j=1}^{d} \mathcal{K}^T(\delta(k, \phi_z(\mathbf{x}_i))) \times w_{jl}}$$

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

– La première étape correspond à des valeurs élevées de T. Si le premier terme est dominant alors la priorité est de préserver la topologie.
– La deuxième étape correspond à des valeurs faibles de T où le deuxième terme est pris en compte dans la fonction de coût. Par conséquent, l'adaptation locale et l'algorithme BiTM converge vers l'algorithme Crouec proposé par Govaert (1983).

3 Expérimentations

Nous avons testé l'algorithme BiTM avec des jeux de données de la base UCI (Frank et Asuncion (2010)). Le tableau 1 indique les paramètres de chaque jeu de données (nombre d'observations, nombre de variables et nombre de classes réelles, ainsi que la taille de la carte utilisée pour l'apprentissage). Afin d'évaluer les performances de BiTM, nous avons utilisé trois indices, la pureté, le rand et le NMI (Normalized Mutual Information) (Strehl et al. (2002)).

Jeux de données	# observations	# variables	Taille de la carte	# classes réelles
isolet5	1559	617	12×12	26
Movement Libras	45	90	5× 5	15
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.*

3.1 Comparaison de BiTM avec les approches de partitionnement

Pour cette première expérimentation, nous comparons les résultats de notre approche BiTM avec les approches suivantes : SOM classique (Kohonen et al. (2001)), HCL (Eisen et al. (1998)), NMF (Paatero et Tapper (1994)) et ONMTF (Long et al. (2005)). Notre objectif à travers cette comparaison est de montrer que BiTM ne modifie pas le comportement général de SOM et fournit des performances comparables aux algorithmes classiques de partitionnement. Nous présentons dans les tables 2, 3, 4 les résultats obtenus en terme d'indices de pureté, rand et NMI.

Le tableau 2 présente les résultats expérimentaux obtenus avec les approches BiTM, SOM, HCL, NMF et ONMTF sur l'indice de pureté. Notre approche BiTM donne des résultats meilleurs ou équivalents sur la plupart des bases de données. Nous remarquons une légère baisse de performance sur les bases : Horse Colic, Cancer Wpbc Ret, glass et isolet5.

En ce qui concerne le tableau 3, BiTM comparé à SOM, HCL, NMF et ONMTF donne

des résultats meilleurs ou équivalents en terme de l'indice de rand sur les bases isolet5, Movement libra, Breast, Sonar mines, Horse colic et heart. Malgré que notre méthode est moins efficace avec les bases Lung cancer, Spectf 1, Cancer Wpbc Ret et glass, BiTM reste stable. Par exemple, avec la base Movement libra, la valeur la plus élevée de l'indice de rand est donnée par SOM 0.943. Notre méthode obtient 0.937. Cependant, HCL obtient une performance de 0.817, NMF 0.789 et ONMTF 0.81. La même analyse peut être faite sur les autres bases de données.

L'analyse des résultats de l'indice NMI présenté dans le tableau 4 montrent que BiTM, SOM et HCL fournissent des performances équivalentes en terme de l'indice NMI. Notre méthode est meilleure ou équivalente à l'approche SOM dans les bases Movement Libras, Breast, Lung cancer, Spectf 1 et Heart. HCL obtient une très forte diminution de l'indice NMI dans la plupart des bases à l'exception de isolet5, Movement Libras, Lung cancer et glass. En dépit d'une faible diminution de l'indice NMI dans certaines bases de données, notre approche fournit des résultats stables pour l'ensemble des bases de données.

dataset	BiTM	SOM	HCL	NMF	ONMTF
isolet5	0.316	**0.433**	0.427	0.107	0.382
MovementLibras	**0.712**	0.711	0.711	0.288	0.311
Breast	**0.978**	0.974	0.633	0.804	0.821
Sonar mines	**0.769**	0.744	0.727	0.601	0.649
LungCancer	**1**	0.906	0.743	0.781	0.875
Spectf 1	**0.759**	0.716	0.65	0.73	0.727
Cancer Wpbc Ret	0.787	**0.828**	0.722	0.762	0.762
HorseColic	0.719	**0.78**	0.713	0.67	0.67
Heart	**0.883**	0.851	0.755	0.62	0.637
glass	0.618	0.623	**0.72**	0.481	0.472

TAB. 2 – *Partitionnement : Résultats de l'indice pureté obtenus avec BiTM , SOM, HCL, NMF et ONMTF.*

dataset	BiTM	SOM	HCL	NMF	ONMTF
isolet5	**0.926**	0.905	0.812	0.471	0.475
MovementLibras	0.937	**0.943**	0.817	0.789	0.81
Breast	**0.687**	0.476	0.499	0.545	0.648
Sonar mines	**0.508**	0.507	0.489	0.504	0.512
LungCancer	0.459	0.425	0.427	0.487	**0.542**
Spectf 1	0.418	0.403	**0.499**	0.436	0.43
Cancer Wpbc Ret	0.435	0.372	**0.54**	0.417	0.408
HorseColic	0.472	0.448	0.449	0.462	**0.488**
Heart	**0.56**	0.529	0.512	0.502	0.506
glass	0.653	**0.752**	0.348	0.689	0.693

TAB. 3 – *Partitionnement : Résultats de l'indice de rand obtenus avec BiTM , SOM, HCL, NMF et ONMTF.*

dataset	BiTM	SOM	HCL	NMF	ONMTF
isolet5	0.439	**0.584**	0.562	0.007	0.015
MovementLibras	**0.811**	0.797	0.555	0.57	0.597
Breast	**0.53**	0.364	0.003	0.193	0.226
Sonar mines	0.158	**0.233**	0.001	0.026	0.047
LungCancer	**0.461**	0.344	0.295	0.111	0.244
Spectf 1	0.1449	**0.185**	0.01	0.025	0.027
Cancer Wpbc Ret	0.081	**0.14**	0.005	0.014	0.017
HorseColic	0.06	**0.128**	0.009	0.03	0.027
Heart	**0.247**	0.225	0.06	0.04	0.036
glass	0.125	**0.463**	0.153	0.231	0.244

TAB. 4 – *Partitionnement : Résultats de l'indice de NMI obtenus avec BiTM , SOM, HCL, NMF et ONMTF.*

3.2 Comparaison de BiTM avec les approches de bi-partitionnement

Afin de comparer BiTM avec les approches de bi-partitionnement, nous avons sélectionné trois approches : CTWC (Getz et al. (2000a)), NBVD (Long et al. (2005)) et CUNMTF (Labiod et Nadif (2011)). Les résultats expérimentaux sont présentés dans les tableaux 5, 6 et 7. Nous signalons que CTWC ne fournit pas de résultats avec la base Movement Libras.

Le tableau 5 résume les résultats expérimentaux de l'indice pureté. Nous remarquons que BiTM fournit les meilleurs résultats sur toutes les bases de données. Dans la plupart des cas, nous constatons une différence remarquable entre les résultats sur l'indice pureté obtenu avec notre méthode et les autres approches. En effet, pour la base Movement libra par exemple, BiTM obtient 0.712, NBVD 0.33 et CUNMTF 0.333. La même constatation pour la base LungCancer où BiTM obtient 1, CTWC 0.718, NBVD 0.875 et CUNMTF 0.843. Nous observons aussi la difficulté d'obtenir de grandes valeurs de l'indice pureté pour la base isolet5.

Comme indiqué dans le tableau 6, BiTM fournit un indice de rand similaire et même meilleur que celui obtenu par les autres méthodes dans la majorité des cas.

Le tableau 7 présente les résultats expérimentaux obtenus avec BiTM, CTWC, NBVD et CUNMTF avec l'indice NMI. Notre approche BiTM fournit les plus hautes valeurs de l'indice NMI pour toute les bases de données excepté pour la base glass.

dataset	BiTM	CTWC	NBVD	CUNMTF
isolet5	**0.316**	0.103	0.073	0.293
MovementLibras	**0.712**	NaN	0.33	0.333
Breast	**0.978**	0.655	0.834	0.834
Sonar mines	**0.769**	0.548	0.644	0.634
LungCancer	**1**	0.718	0.875	0.843
Spectf 1	**0.759**	0.727	0.727	0.727
Cancer Wpbc Ret	**0.787**	0.762	0.762	0.762
HorseColic	**0.719**	0.67	0.67	0.673
Heart	**0.883**	0.555	0.674	0.674
glass	**0.618**	0.523	0.462	0.462

TAB. 5 – *Bi-partitionnement : Comparaison en utilisant l'indice de pureté obtenu avec BiTM, CTWC, NBVD et CUNMTF.*

dataset	BiTM	CTWC	NBVD	CUNMTF
isolet5	**0.926**	0.91	0.502	0.508
MovementLibras	**0.937**	NaN	0.845	0.84
Breast	0.687	0.505	0.659	**0.688**
Sonar mines	**0.508**	0.502	0.514	**0.508**
LungCancer	0.459	**0.556**	**0.556**	0.536
Spectf 1	0.418	**0.513**	0.42	0.42
Cancer Wpbc Ret	0.435	**0.524**	0.414	0.417
HorseColic	**0.472**	0.463	0.46	0.459
Heart	**0.56**	0.498	0.513	0.515
glass	0.653	0.69	**0.693**	0.69

TAB. 6 – *Bi-partitionnement : Comparaison en utilisant l'indice de Rand obtenu avec BiTM, CTWC, NBVD et CUNMTF.*

dataset	BiTM	CTWC	NBVD	CUNMTF
isolet5	**0.439**	0.077	0.137	0.186
MovementLibras	**0.811**	NaN	0.688	0.667
Breast	**0.53**	0.01	0.243	0.0233
Sonar mines	**0.158**	0.006	0.057	0.04
LungCancer	**0.461**	0.041	0.309	0.261
Spectf 1	**0.1449**	0.001	0.016	0.016
Cancer Wpbc Ret	**0.081**	0.034	0.024	0.031
HorseColic	**0.06**	0.003	0.03	0.028
Heart	**0.247**	0.001	0.063	0.061
glass	0.125	**0.38**	0.246	0.245

TAB. 7 – *Bi-partitionnement : Comparaison en utilisant l'indice NMI obtenu avec BiTM, CTWC, NBVD et CUNMTF.*

3.3 Visualisation

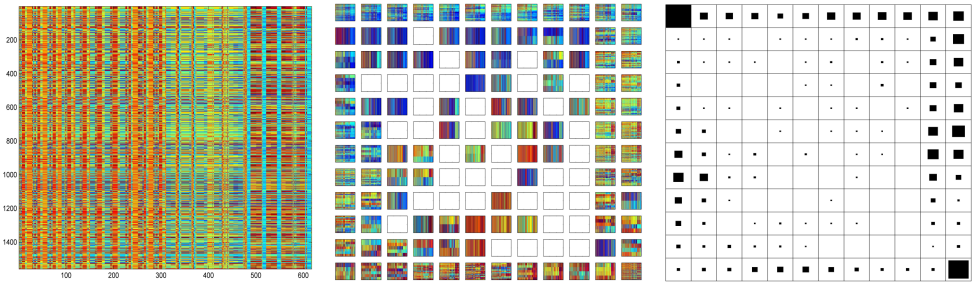
Dans cette section nous montrons l'apport visuel de l'approche proposée. Il est clair que BiTM se base sur les visualisations intuitives des cartes auto-organisatrices. Les figures 1(a), 1(b) et 1(c) sont des cartes de BiTM obtenues à partir du jeu de données isolet5. La figure 1(a) est dédiée à la visualisation de la base de données organisées en fonction des groupes de lignes et de colonnes. Cette figure peut être obtenue par toute méthode 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 de BiTM. Ainsi, chaque cellule de la carte est associée à la partition des observations et des variables. Cette organisation est illustrée par la figure 1(b). Par exemple, les groupes en haut à gauche de la carte ont des valeurs faibles représentées par des couleurs bleues. Tandis que ceux avec de fortes valeurs (au milieu en bas) sont représentés par des couleurs plus vives (rouge). La figure 1(c) indique la cardinalité de chaque cellule. Les cellules sont représentées par un carré dont la taille varie proportionnellement avec le nombre d'observations associées.

Il est également possible de zoomer sur chaque cellule de la carte pour analyser l'organisation des observations et des variables dans chaque cellule. Les résultats obtenus en zoomant sur la carte 1(b) sont représentés dans la figure 2. Comme nous utilisons la notion du voisinage dans le modèle BiTM, nous constatons que la couleur est relativement similaire lorsque les variables sont proches.

Finalement 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.

4 Conclusion et perspectives

Nous avons constaté après l'étude comparative avec des méthodes de partitionnement et de bi-partitionnement que BiTM est une méthode de bi-partitionnement efficace. La princi-

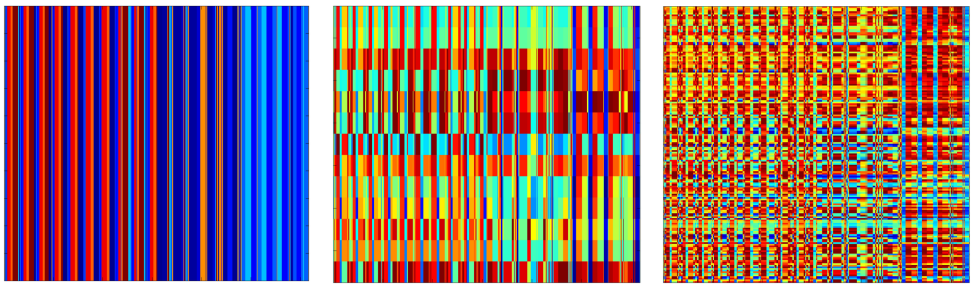


(a) La base de données organisées en fonction de l'ordre des observations et des variables de la classification croisée.

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

(c) Cardinalité de la carte BiTM.

FIG. 1 – *Visualisation de la base de données isolet5 en utilisant BiTM. Chaque cellule dans la figure 1(b) et 1(c) indique une cellule de la carte.*



(a) zoom sur la cellule # 2.

(b) zoom sur la cellule # 48.

(c) zoom sur la cellule # 144.

FIG. 2 – *Figure obtenu en zoomant sur quelques cellules dans la carte BiTM présentée dans la figure 1(b).*

pale nouveauté du BiTM est l'utilisation d'un modèle topologique pour organiser la matrice des données en blocs homogènes, tout en prenant en compte simultanément les lignes et les colonnes. La série d'expériences que nous avons réalisées nous ont permis de valider notre méthode et d'analyser ses performances à partir de nombreux critères. Ces résultats expérimentaux démontrent que notre algorithme identifie les bi-clusters et a de bonnes performances par rapport à certains algorithmes de classification croisée. Nombreuses sont les perspectives qu'offre notre approche telle que l'amélioration de l'utilité de BiTM en l'adaptant à des données binaires et mixtes.

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.

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

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*.

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

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

Eisen, M., P. Spellman, P. Brown, et D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns.

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

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*.

Hartigan, J. A. (1972). Direct Clustering of a Data Matrix. *Journal of the American Statistical Association 67*(337), 123–129.

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.

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.

Paatero, P. et U. Tapper (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. pp. 111–126.

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

Shang, F., L. C. Jiao, et F. Wang (2012). Graph dual regularization non-negative matrix factorization for co-clustering. *Pattern Recogn. 45*(6), 2237–2250.

Strehl, A., J. Ghosh, et C. Cardie (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research 3*, 583–617.

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.

Yoo, J. et S. Choi (2010). Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Inf. Process. Manage. 46*(5), 559–570.

Summary

In this paper, we propose a new bi-clustering algorithm based on self-organizing maps titled BiTM (Bi-clustering using Topological Map). BiTM provides a simultaneous clustering of rows and columns of the data matrix in order to increase the homogeneity of bi-clusters by respecting neighborhood relationship and using a single map. BiTM maps provide a new topological visualization of the bi-clusters. Experimental results and comparison studies show that BiTM improves the results in term of bi-clustering and visualization.
Keywords : Bi-clustering, co-clustering, self-organizing maps.