Binary Block GTM : Carte auto-organisatrice probabiliste pour les grands tableaux binaires

Rodolphe Priam*, Mohamed Nadif**, Gérard Govaert***

*LMA Poitiers, UMR CNRS 6086, Université de Poitiers,
BP 30179, 86962 Futuroscope Chasseneuil Cedex, France
rpriam@gmail.com
**CRIP5, Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris, France
mohamed.nadif@univ-paris5.fr
***Heudiasyc, UMR CNRS 6599, Université de Technologie de Compiègne,
BP 20529, 60205 Compiègne Cedex, France
gerard.govaert@utc.fr

Résumé. Ce papier présente un modèle génératif et son estimation permettant la visualisation de données binaires. Notre approche est basée sur un modèle de mélange de lois de Bernoulli par blocs et les cartes de Kohonen probabilistes. La méthode obtenue se montre à la fois parcimonieuse et pertinente en pratique.

1 Introduction

Bien que les méthodes d'analyse factorielle soient très puissantes et contribuent efficacement à la visualisation des données, les grands échantillons nécessitent de nouvelles méthodes mieux adaptées. En effet, les algorithmes de décomposition matricielle rencontrent leurs limites sur les grands tableaux numériques ; en outre, la construction de nombreux plans de projection, du fait des grandes dimensions, rend la tâche d'interprétation difficile pour recouper les informations disséminées sur ces plans. Finalement une grande quantité de données implique une grande quantité d'informations à synthétiser et des relations complexes entre individus et/ou variables étudiés. Il est alors possible, dans ce contexte, d'utiliser les cartes de Kohonen ou cartes auto-organisatrices (SOM) (Kohonen, 1997) qui sont des méthodes de classification automatique utilisant une contrainte de voisinage sur les classes pour conférer un sens topologique aux partitions obtenues. La carte auto-organisatrice originelle peut être vue comme une variante de l'algorithme des *k-means* (MacQueen, 1967) intégrant une contrainte d'ordre topologique sur les centres.

Lorsque la matrice des données $\mathbf{x}$ est définie sur un ensemble I d'objets (lignes, observations) et un ensemble J de variables (colonnes, attributs), différentes approches de classification automatique sont utilisées et la plupart des algorithmes proposés concerne généralement un des deux ensembles. Ces algorithmes peuvent être modélisés par différentes approches. Celle qui a suscité le plus d'intérêt ces dernières années est incontestablement l'approche modèle de mélange (McLachlan et Peel, 2000). Dans ce cadre, il a été proposé diverses versions probabilistes de SOM telles que dans (Lebbah et al., 2007; Verbeek et al., 2005; Luttrell, 1994).

Le papier (Govaert et Nadif, 2003) présente une extension du modèle de mélange pour répondre à l'objectif de la classification croisée dite aussi classification par blocs qui permet de tenir compte de I et J simultanément. Différents modèles ont été proposés pour tenir compte de chaque type de données.

Ces méthodes (Dhillon, 2001) ont un grand intérêt en *data mining* car elles sont particulièrement appropriées pour les grands ensembles de données en grande dimension. Elles ne sont pourtant pas encore employées en visualisation alors qu'elles ont le potentiel pour fournir un outil très efficace et parcimonieux. En effet, elles utilisent beaucoup moins de paramètres que les modèles connus usuels tels que les modèles classiques de mélange.

Pour analyser le contenu d'un ensemble de données, la visualisation est une étape crucial pour laquelle les modèles génératifs sont devenus très utiles. En effet, la taille croissante des ensembles de données rencontrés permet une estimation pertinente de variables cachées synthétisant de manière interprétable l'information contenue dans les données. Pour toutes ces raisons, nous proposons dans ce papier de traiter la question de la visualisation par une approche basée sur le modèle de mélange croisé parcimonieux et l'algorithme GTM (Bishop et al., 1998), méthode de auto-organisatrice probabiliste basée sur un modèle gaussien.

Ce papier est organisé comme suit. Le deuxième paragraphe présente une brève introduction du modèle de mélange croisé et une description rapide de l'algorithme *Block EM*. Le troisième paragraphe est consacré au développement, dans le cas binaire, de l'algorithme *Block Generative Topographic Model* ou *Block GTM*. Cet algorithme peut être vu comme une extension efficace du GTM à un modèle de mélanges de Bernoulli par blocs. Un algorithme d'estimation y est présenté. Le quatrième paragraphe présente des expériences numériques à partir de deux matrices binaires textuelles. Enfin, le dernier paragraphe résume les principaux résultats du papier et les perspectives originales de cette approche.

Dans la suite, la matrice de données est notée $\mathbf{x} = \{(x_{ij}); i \in I \text{ et } \in J\}$, où I est un ensemble de n objets (lignes, observations) et J est un ensemble de d variables (colonnes, attributs). Une partition $\mathbf{z}$ en g classes de l'échantillon I sera représentée par la matrice de classification $(z_{ik}\,;\, i = 1, \ldots, n\,;\, k = 1, \ldots, g)$ où $z_{ik} = 1$ si i appartient à la classe k et 0 sinon. Une notation similaire sera utilisée pour la partition $\mathbf{w}$ en m classes de l'ensemble J. Par souci de simplification des formules, les intervalles de variation des indices ne seront pas spécifiés, par exemple, nous noterons $\sum_{i,j,k,\ell}$ au lieu $\sum_{i=1}^{n}\sum_{j=1}^{d}\sum_{k=1}^{g}\sum_{\ell=1}^{m}$.

2 L'algorithme *Block EM*

L'objectif de la classification par blocs est d'essayer de résumer cette matrice par des blocs homogènes. Le problème peut être étudié sous l'approche d'une partition simultanée des deux ensembles I et J en g et m classes respectivement. Dans (Govaert, 1983, 1995) plusieurs algorithmes ont été proposés pour obtenir une classification par blocs sur des tableaux de contingence ou plus généralement sur des tables qui ont les même propriétés : des données binaires, continues ou catégorielles.

Dans (Govaert et Nadif, 2003), ces méthodes ont été modélisées par une approche basée sur des mélanges de lois. Dans le contexte du problème de la classification par blocs, la formulation du modèle de mélange classique peut être étendue pour proposer un modèle en blocs latents

défini par une distribution en sommant sur l'ensemble des affectations de $I \times J$:

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} p(\mathbf{z};\boldsymbol{\theta})p(\mathbf{w};\boldsymbol{\theta})f(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta})$$

où $\mathcal{Z}$ et $\mathcal{W}$ dénote les ensembles de toutes les affectations possibles $\mathbf{z}$ de I et $\mathbf{w}$ de J. Comme pour l'analyse en classes latentes, les $n \times d$ variables aléatoires X_{ij} générant les cellules x_{ij} observées sont supposées être indépendantes lorsque $\mathbf{z}$ et $\mathbf{w}$ sont fixés ; nous avons alors

$$f(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}$$

où $\varphi(.,\alpha_{k\ell})$ est une distribution définie sur l'ensemble des réels $\mathbb{R}$.

Par exemple, lorsque les données sont binaires, en notant $\boldsymbol{\theta} = (\mathbf{p},\mathbf{q},\alpha_{11},\ldots,\alpha_{gm})$, où $\mathbf{p} = (p_1,\ldots,p_g)$ et $\mathbf{q} = (q_1,\ldots,q_m)$ sont les vecteurs de probabilités p_k et q_ℓ qu'une ligne et une colonne appartienne au k$^{\text{e}}$composant et au ℓ^{e}composant respectivement, nous obtenons le modèle par blocs latents de Bernoulli défini par la distribution suivante :

$$\varphi(x_{ij};\alpha_{k\ell}) = (\alpha_{k\ell})^{x_{ij}}(1-\alpha_{k\ell})^{1-x_{ij}}.$$

Utiliser ce modèle est nettement plus parcimonieux qu'utiliser un modèle classique de mélange sur chaque ensemble I et J. Par exemple, avec $n = 1000$ objets et $d = 500$ variables et des probabilités de classes égales $p_k = 1/g$ et $q_\ell = 1/m$, si on a besoin de faire la classification automatique d'une matrice binaire en $g = 4$ classes en lignes et $m = 3$ classes en colonnes, le modèle par blocs latents de Bernoulli impliquera l'estimation de 12 paramètres $(\alpha_{k\ell}, k = 1,\ldots,4, \ell = 1,\ldots,3)$ au lieu de $(4 \times 500 + 3 \times 1000)$ pour les deux modèles de mélange de Bernoulli appliqués à I et J séparément.

Maintenant nous nous intéressons à l'estimation d'une valeur optimale de $\boldsymbol{\theta}$ par l'approche du maximum de vraisemblance associé à ce modèle de mélange par blocs. Pour ce modèle, les données complétées sont le vecteur $(\mathbf{x},\mathbf{z},\mathbf{w})$ où les vecteurs non observés $\mathbf{z}$ et $\mathbf{w}$ sont les labels. La vraisemblance classifiante $L(\boldsymbol{\theta};\mathbf{x},\mathbf{z},\mathbf{w}) = \log f(\mathbf{x},\mathbf{z},\mathbf{w};\boldsymbol{\theta})$ est notée $L_C(\mathbf{z},\mathbf{w},\boldsymbol{\theta})$. L'algorithme EM (Dempster et al., 1977) maximise la vraisemblance $L_M(\boldsymbol{\theta})$ par rapport à $\boldsymbol{\theta}$ itérativement en maximisant l'espérance conditionnelle de la vraisemblance des données complétées $L_C(\mathbf{z},\mathbf{w},\boldsymbol{\theta})$ par rapport à $\boldsymbol{\theta}$, étant donné une estimation précédente courante $\boldsymbol{\theta}^{(t)}$ et les données observées $\mathbf{x}$:

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(t)}) = \sum_{i,k} c_{ik}^{(t)} \log p_k + \sum_{j,\ell} d_{j\ell}^{(t)} \log q_\ell + \sum_{i,j,k,\ell} e_{ikj\ell}^{(t)} \log \varphi(x_{ij};\alpha_{k\ell})$$

où $c_{ik}^{(t)}$, $d_{j\ell}^{(t)}$, et $e_{ikj\ell}^{(t)}$ sont respectivement les probabilités a postériori sur les lignes, les colonnes, et les cellules, à l'itération t.

Malheureusement, la structure de dépendance des variables X_{ij} du modèle entraine des difficultés pour la détermination de $e_{ikj\ell}^{(t)}$. Pour résoudre ce problème, une approximation variationnelle remplaçant $e_{ikj\ell}^{(t)}$ par le produit $c_{ik}^{(t)} d_{j\ell}^{(t)}$ permet de fournir une bonne solution (Govaert et Nadif, 2005).

Dans la section suivante, nous développons un algorithme d'apprentissage intégrant une contrainte d'ordre topologique sur les paramètres $\alpha_{k\ell}$.

3 Modèle et estimation

Nous présentons le *Binary Bock GTM*, une carte auto-organisatrice générative par blocs pour une matrice binaire $\mathbf{x}$ dont chaque cellule x_{ij} est un réel 0 ou 1. Pour induire une auto-organisation topologique des densités gaussiennes, une approche par le GTM considère des coordonnées $2d$ pour les noeuds d'une grille rectangulaire imaginaire qui représente l'espace de projection. Ce graphe planaire peut être vu comme une discrétisation d'une partie du plan sur lequel les données, les n lignes de la matrice, vont être projetées. Comme chaque noeud doit correspondre à une classe, chaque point 2d sur le plan est alors sujet à une transformation non linéaire afin d'être amené dans un espace de dimension h supérieure. Une projection linéaire permet alors d'obtenir des centres de même dimension que les individus vectoriels.

Plus formellement, afin d'obtenir une auto-organisation des probabilités $\alpha_{k\ell}$, ces dernières sont paramétrées par les g coordonnées s_k dessinant une grille rectangulaire régulière sur le plan. Ces coordonnées sont projetées dans un espace de plus grande dimension h, soit en prenant pour les applications ϕ des bases fonctionnelles de type noyau, $\xi_k = \Phi(s_k) = (\phi_1(s_k), \phi_2(s_k), \cdots, \phi_h(s_k))$, avec par exemple, $\phi(s_k) = exp\Big(- \frac{||s_k - m_\phi||^2}{\sigma_\phi^2} \Big)$ où m_ϕ est un centre posé ad'hoc et σ_ϕ une variance bien choisie. Finalement, la paramétrisation nécessite l'estimation de m vecteurs de dimension h inconnus nommés w_ℓ. On écrit les probabilités du BEM binaire original à l'aide de fonctions sigmoïdes $\alpha_{k\ell} = \sigma(w_\ell^T \xi_k)$ où $\sigma(y) = e^y/(1+e^y)$, comme montré en figure 1. Le vecteur de paramètres devient $\boldsymbol{\theta} = (\mathbf{p}, \mathbf{q}, w_1, w_2, \cdots, w_m)$.

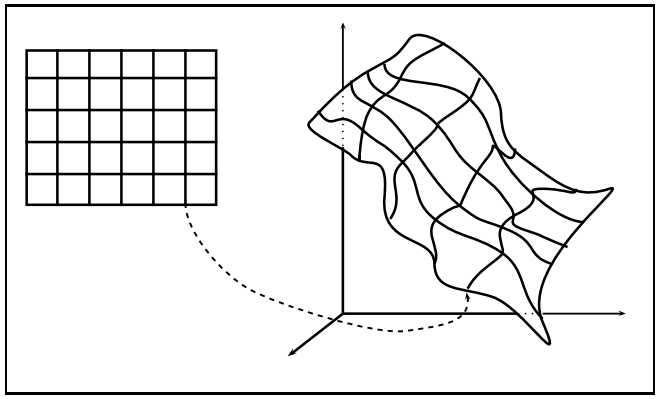

FIG. 1 – *A gauche, la grille rectangulaire des s_k, sur la droite, l'espace des distributions φ. Le graphique représente la paramétrisation non linéaire des sigmoïdes. Chaque coordonnée s_k, pour $k = 1, \cdots, g$, de la grille est transformée de façon non linéaire pour se retrouver dans l'espace des distributions multivariées de Bernoulli par la transformation $\sigma(w_\ell^T \xi_k)$, pour $l = 1, \cdots, m$.*

La matrice $g \times m$ de probabilités est remplacée par la matrice $h \times m$, et le modèle demeure parcimonieux puisque h est petit en pratique, quelques dizaines. Donc, en reprenant à nouveau l'exemple d'une matrice binaire de 1000 lignes et 500 colonnes de la section précédente, nous aboutissons à environ 100 paramètres, compte tenu du choix à effectuer sur la valeur de h, qui est toujours peu comparativement à une approche de mélange classique. Ensuite, les para-

mètres inconnus sont estimés en trouvant un maximum local de la log-vraisemblance par un algorithme EM (Dempster et al., 1977).

Grâce à l'approximation variationnelle de $e_{ikj\ell}^{(t)}$, il peut être montré (Govaert et Nadif, 2005) que la maximisation de la log-vraisemblance du *Block EM* est réalisée en maximisant alternativement deux critères conditionnels $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}|\mathbf{c})$ et $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}|\mathbf{d})$ avec $\mathbf{c} = (c_{ik}; i = 1, \ldots, n; k = 1, \ldots, g)$ et $\mathbf{d} = (d_{j\ell}; j = 1, \ldots d; \ell = 1, \ldots, m)$. A la convergence en une position stable de $\boldsymbol{\theta}$, le paramètre optimal est nommé $\hat{\boldsymbol{\theta}}$. Ici, considérant des paramètres w_ℓ, ces deux critères prennent la forme suivante :

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}|\mathbf{d}) = \sum_{i,k} c_{ik}^{(t)} \Big\{ \sum_\ell u_{i\ell}^{(t)} w_\ell^T \xi_k - d_\ell^{(t)} \log(1 + e^{1+w_\ell^T \xi_k}) \Big\}$$

avec $u_{i\ell}^{(t)} = \sum_j d_{j\ell}^{(t)} x_{ij}$, $d_\ell^{(t)} = \sum_j d_{j\ell}^{(t)}$, et un critère similaire $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}|\mathbf{c})$ pour les colonnes avec ${v_{jk}}^{(t)} = \sum_i c_{ik}^{(t)} x_{ij}$ et $c_k^{(t)} = \sum_i c_{ik}^{(t)}$. La maximisation de ces deux espérances, effectuée à l'aide de la méthode du gradient, conduit aux relations suivantes :

$$w^{(t+\frac{1}{2})} = argmax_w \; Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}|\mathbf{d}) \quad \text{et ensuite} \quad w^{(t+1)} = argmax_w \; Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t+\frac{1}{2})}|\mathbf{c}).$$

En dérivant les deux critères, on obtient les vecteurs de gradient $\mathbf{Q}_u^{(t)}$, $\mathbf{Q}_v^{(t)}$, et les matrices hessiennes $\mathbf{H}_u^{(t)}$, $\mathbf{H}_v^{(t)}$. Comme les hessiennes sont diagonales par blocs, la log-vraisemblance est augmentée à chaque pas EM par deux pas de montée de type Newton-Raphson, pour ℓ de 1 à m, ce qui correspond à un algorithme EM généralisé.
En posant, $\Phi = (\xi_1^T, \xi_2^T, \cdots, \xi_g^T)^T$ la $g \times h$ matrice des bases fonctionnelles, nous obtenons :

$$\begin{aligned} w_\ell^{(t+\frac{1}{2})} &= w_\ell^{(t)} + \tfrac{1}{d_{(\ell)}} \Big(\Phi^T G F_\ell \Phi \Big)^{-1} \Big(\Phi^T C u_\ell - d_{(\ell)} \, \Phi^T G \alpha_\ell \Big) \\ w_\ell^{(t+1)} &= w_\ell^{(t+\frac{1}{2})} + \tfrac{1}{d_{(\ell)}} \Big(\Phi^T G F_\ell \Phi \Big)^{-1} \Big(\Phi^T V d_\ell - d_{(\ell)} \, \Phi^T G \alpha_\ell \Big) \end{aligned}$$

où C est la matrice $g \times n$ des probabilités a postériori avec $c_{ik}^{(t)}$ pour cellules, V la matrice $g \times d$ avec $v_{jk}^{(t)}$ pour cellules, G la matrice diagonale $g \times g$ avec $c_k^{(t)}$ sur sa diagonale, F_ℓ la matrice diagonale $g \times g$ avec $\alpha_{k\ell}^{(t)}(1 - \alpha_{k\ell}^{(t)})$ à ℓ fixé sur sa diagonale, α_ℓ le vecteur $g \times 1$ avec les $\alpha_{k\ell}^{(t)}$ à ℓ fixé pour valeurs, u_ℓ le vecteur $n \times 1$ avec les $u_{i\ell}^{(t)}$ à ℓ fixé pour composantes, d_ℓ le vecteur $d \times 1$ composé des $d_{j\ell}^{(t)}$ à ℓ fixé, et $d_{(\ell)} = d_\ell^{(t)}$. Enfin pour $1 \leq \ell \leq m$, on a $w_\ell^{(t)} \in \mathbb{R}^h$.

En itérant t et le calcul de $w_\ell^{(t+1)}$, les valeurs consécutives courantes convergent vers un maximum d'une approximation de $L_M(\boldsymbol{\theta})$. Un biais bayésien (Bishop et al., 1998) peut éventuellement être ajouté pour améliorer la stabilité numérique des estimations. La forme matricielle obtenue par une approche de gradient du second ordre est analogue à une étape d'IRLS (McCullagh et Nelder, 1983). Une alternative serait un gradient au premier ordre sous optimal en pratique. On remarque enfin que la symétrie des formules du BEM originale est ici absente du fait que seules les lignes sont projetées par la méthode proposée.

4 Expériences numériques

Nous évaluons notre nouvelle méthode de projection à partir de deux matrice binaires de données textuelles. Les paramètres utilisés dans nos expériences sont $m = 10$, $g = 81$ et

$h = 28$ pour les deux bases de textes.

La projection sur le plan d'un échantillon de données binaires par le modèle *Block GTM* peut s'effectuer de diverses manières, dont essentiellement :
- La représentation matricielle qui place en s_{k^*} l'ensemble des individus affectés à la classe associée au noeud, tels que $\hat{z}_i = k^*$. Cette affectation obéit à la règle du maximum a postériori (MAP) donc $\hat{z}_i = argmax_k\ \hat{c}_{ik}$. Dans le cas du SOM, l'affectation utilise la distance euclidienne entre le vecteur centre et le vecteur donnée.
- La deuxième représentation, que nous avons utilisée dans la suite car celle-ci est plus fidèle à la classification floue obtenue, consiste en une projection par position moyenne sur le plan :

$$\hat{p}_i = \sum_k \hat{c}_{ik} s_k$$

On remarque que la projection MAP correspond à la projection moyenne dans laquelle on remplace la matrice de classification floue de cellules $(\hat{c}_{ik})$ par la matrice de classification dure de cellules $(\hat{z}_{ik})$.

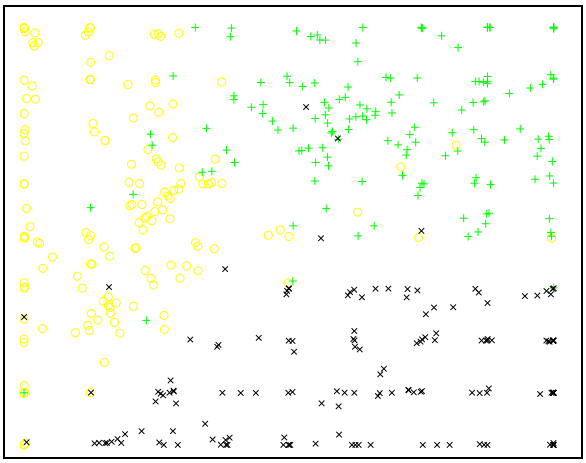

FIG. 2 – *Projection par Binary Block GTM de la matrice textuelle* 449×167 *des données Classic 3.*

La première matrice est projetée pour tester le modèle avec trois classes. Ces données correspondent à un échantillon de la matrice Classic 3 (Dhillon et al., 2003), qui est constituée de trois bases d'articles scientifiques : $Medline$, $Cisi$, $Cranfield$. Par tirage au hasard, 450 documents ont été sélectionnés avec 150 documents dans chaque classe. Seuls les mots les plus fréquents (au dessus du seuil 30) ont été retenus. La matrice finale est de 449 lignes et 167 colonnes. La projection sur la figure 2 sépare les classes sans erreur quasiment. Les $outliers$ peuvent s'expliquer par le fait que le tableau original est de contingence, et également que les classes ne sont pas exactement disjointes comme le révèle les *benchmarks* relatifs.

La seconde matrice textuelle présente quatre classes et compte 400 documents décrits par 100 termes (Girolami, 2001). Le vocabulaire a été choisi par tri selon l'information mutuelle

évaluée grâce aux labels des classes. La projection de ces textes, à la figure 3, révèle 4 *clusters* facilement reconnaissables et correspondant aux quatre groupes de discussion "sci.crypt", "sci.space,"sci.med", et "soc.religion.christian"; dans chacun, 100 *news* ont été tirés au hasard. Les classes sont bien séparées avec des frontières précises et les classes de mots obtenues peuvent être interprétées. Notons que la carte obtenue avec notre approche est assez similaire à la carte auto-organisatrice probabiliste basée sur un modèle multinomial asymétrique (Kabán et Girolami, 2001), qui est moins parcimonieux.

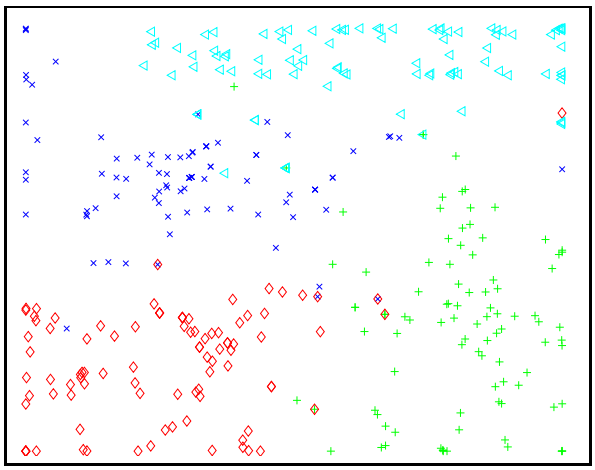

FIG. 3 – *Projection par Binary Block GTM de la matrice textuelle* 400×100 *des données newsgroups.*

5 Conclusion et perspectives

Nous avons proposé une carte auto-organisatrice probabiliste. Celle-ci est obtenue par l'utilisation d'un modèle de mélange de Bernoulli croisé et de l'algorithme GTM. Notre méthode, appelée *Binary Block GTM*, est efficace et parcimonieuse. En effet, le nombre de paramètres inconnus est égal à $h \times m$, ce qui est très peu comparativement à un modèle contraint de mélange de lois de Bernoulli (Girolami, 2001) ou une approche dyadique comme un pLSA binaire contraint (Priam et Nadif, 2006). Quelle que soit g la taille de la carte de projection, quel que soit n le nombre de lignes, le nombre de paramètres du modèle ne croît qu'avec le nombre de classes en colonnes. En conclusion, le modèle présenté apparait clairement comme un excellent candidat pour s'attaquer aux problèmes du *data mining*. Il serait intéressant d'étendre cette approche au tableau de contingence en proposant un *Block* GTM adapté.

Références

Bishop, C. M., M. Svensén, et C. K. I. Williams (1998). Developpements of generative topographic mapping. *Neurocomputing 21*, 203–224.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum-likelihood from incomplete data via the em algorithm. *J. Royal Statist. Soc. Ser. B., 39*.

Dhillon, I. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Seventh ACM SIGKDD Conference*, San Francisco, California, USA, pp. 269–274.

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

Girolami, M. (2001). The topographic organization and visualization of binary data using multivariate-bernoulli latent variable models. *IEEE Transactions on Neural Networks 20*(6), 1367–1374.

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

Govaert, G. (1995). Simultaneous clustering of rows and columns. *Control and Cybernetics 24*(4), 437–458.

Govaert, G. et M. Nadif (2003). Clustering with block mixture models. *Pattern Recognition 36*, 463–473.

Govaert, G. et M. Nadif (2005). An EM Algorithm for the Block Mixture Model. *IEEE Trans. Pattern Anal. Mach. Intell. 27*(4), 643–647.

Kabán, A. et M. Girolami (2001). A combined latent class and trait model for analysis and visualisation of discrete data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 859–872.

Kohonen, T. (1997). *Self-organizing maps*. Springer.

Lebbah, M., N. Rogovschi, et Y. Bennani (2007). Besom : Bernoulli on self organizing map. In *International Joint Conferences on Neural Networks, IJCNN'2007*.

Luttrell, S. P. (1994). A Bayesian analysis of self-organising maps. *Neural Computation 6*, 767–794.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. L. Cam et J. Neyman (Eds.), *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 281–297. University of California Press.

McCullagh, P. et J. Nelder (1983). *Generalized linear models.* London : Chapman and Hall.

McLachlan, G. J. et D. Peel (2000). *Finite Mixture Models*. New York : John Wiley and Sons.

Priam, R. et M. Nadif (2006). Carte auto-organisatrice probabiliste sur données binaires (in french). *RNTI (EGC'2006 proceedings)*, 445–456.

Verbeek, J. J., N. A. Vlassis, et B. J. A. Kröse (2005). Self-organizing mixture models. *Neurocomputing 63*, 99–123.

Summary

This article presents a generative model and its estimation allowing to visualize binary data. Our approach is based on the Bernoulli block mixture model and the probabilistic self-organizing maps. The obtained method is parcimonious and relevant on real data.