Visualisation et classification avec les cartes topologiques catégorielles

Mustapha Lebbah[*,**], Fouad Badran[**],
Sylvie Thiria[*]

[*] Laboratoire LODYC, Université Paris 6, Tour 45-5[e]étage, boite 100
4 place Jussieu 75252 Paris cedex 05 France
lebbah,thiria@lodyc.jussieu.fr,
[**] CEDRIC, Conservatoire National des Arts et Métiers,
292 rue Saint Martin, 75003 Paris, France
badran@cnam.fr

Résumé. Ce papier introduit les cartes topologiques dédiée à la visualisation et à la classification de données composées de variables catégorielles. Pour visualiser ou classer ces données par des cartes topologiques, les méthodes classiques utilisent une phase de codage de prétraitement de ces données en données numériques et appliquent l'algorithme classique des cartes topologiques. Dans ce papier nous proposons un modèle de cartes topologiques dédiées aux données catégorielles. Ce modèle est basé sur un formalisme probabiliste où chaque cellule est représentée par une table de probabilités. Deux exemples réels permettent de valider ce modèle. Les résultats obtenus montrent l'apport de ce modèle dans la visualisation et la classification de données catégorielles.

1 Introduction

La visualisation des données est une étape importante dans la phase exploratoire d'une analyse de données. Cette étape est plus difficile quand il s'agit de données qualitatives pour lesquelles il existe moins de méthodes standard. Les cartes topologiques sont de plus en plus utilisées comme outils de visualisation, puisqu'elles permettent de projeter sur des espaces discrets qui sont généralement de dimensions deux. Le modèle de base, proposé par Kohonen [9], est dédié uniquement aux données numériques, il a cependant été appliqué avec succès au traitement de donnés textuelles [8]. Cet algorithme a été appliqué aussi sur des données binaires (données catégorielles codées en binaires) précédé d'un prétraitement ou d'un changement de représentation spécifique des données [7, 13]. Des extensions et reformulations du modèle de Kohonen ont été proposé dans la littérature : Cartes topologiques probabilistes [1] , Generative Topographic Mapping [6, 2]. En se basant sur le formalisme classique des cartes topologiques, nous avons déjà proposé un modèle de cartes topologiques dédiés aux données binaires [10]. Dans ce papier, nous présentons le modèle CTM de cartes topologiques dédié aux données catégorielles. Ce modèle est basé sur le formalisme des cartes topologiques probabilistes, l'apprentissage consiste alors à estimer les paramètres du modèle en maximisant la fonction de vraisemblance des observations de la base d'apprentissage. L'algorithme d'apprentissage que nous proposons est une application de l'algorithme EM. Au paragraphe 2 nous présentons le modèle CTM et l'algorithme d'apprentissage associé,

ainsi que l'utilisation de ce modèle comme classifieur. Nous présentons au 3ième paragraphe deux applications. La première porte sur une enquête semiologie, elle permet de montrer l'apport de cet algorithme pour la visualisation des données catégorielles, la seconde application porte sur une enquête d'assurance qui permet de montrer l'apport de cet algorithme pour la classification. En fin, nous concluons en donnant quelques perspectives.

2 Carte Topologique Catégorielle

2.1 Modèle CTM

Comme tout modèle de cartes topologiques nous supposons que l'on dispose d'une carte discrète $\mathcal{C}$ ayant N_{cell} cellules structurées par un graphe non orienté. Cette structure de graphe permet de définir une distance, $d(r,c)$ entre deux cellules r et c de $\mathcal{C}$, comme étant la longueur de la plus courte chaîne permettant de relier les cellules r et c. Le modèle CTM est basé sur le formalisme probabiliste des cartes topologiques, qui associe à chaque cellule c de $\mathcal{C}$ une probabilité $p_c(\mathbf{z})$ où $\mathbf{z}$ est un vecteur de l'espace des données. En suivant le formalisme bayésien, introduit dans Luttrel [14], chaque cellule de $\mathcal{C}$ modélise un mélange de lois de probabilités. Pour simplifier la compréhension de la modélisation, on fait comme si la carte $\mathcal{C}$ était dupliquée en deux cartes identiques à $\mathcal{C}$: $\mathcal{C}_1$ et $\mathcal{C}_2$ (dans la pratique il y'en a qu'une seule carte). Ce formalisme suppose que les observations $\mathbf{z}$ sont générées de la manière suivante : on commence par choisir une cellule c_2 de $\mathcal{C}_2$ suivant les probabilités a priori $p(c_2)$, celle-ci est alors propagée à la seconde carte $\mathcal{C}_1$ en suivant la probabilité conditionnelle $p(c_1/c_2)$ et enfin $\mathbf{z}$ est générée suivant la distribution de probabilité $p(\mathbf{z}/c_1)$. Ce formalisme nous amène à définir la distribution de probabilité $p(\mathbf{z})$ par un mélange de probabilités : $p(\mathbf{z}) = \sum_{c_2 \in \mathcal{C}_2} p(c_2) p_{c_2}(\mathbf{z})$, avec $p_{c_2}(\mathbf{z}) = \sum_{c_1 \in \mathcal{C}_1} p(c_1/c_2) p(\mathbf{z}/c_1)$, où la probabilité conditionnelle $p(c_1/c_2)$ est supposée connue.

Dans la suite et afin d'introduire la notion de voisinage sur la carte, nous supposons que $p(c_1/c_2)$ est égale à $\frac{K^T(\delta(c_1,c_2))}{\sum_{r \in C_1} K^T(\delta(r,c_2))}$, où K^T est une fonction de voisinage dépendant du paramètre T et qui peut s'exprimer par $K^T(d) = K(d/T)$, étant K une fonction noyau particulière positive et symétrique (avec $\lim_{|x| \to \infty} K(x) = 0$).

On suppose par la suite que les données $\mathbf{z}$ sont des vecteurs à n composantes catégorielles $\mathbf{z} = (z^1, z^2, ..., z^k, ..., z^n)$, où chaque composante $z^k \in M_k$ qui est l'ensemble fini formé par l'énumération des m_k modalités $\{x_1^k, x_2^k, ..., x_{m_k}^k\}$ de la k^{ieme} composante de $\mathbf{z}$, dans ce cas $\mathbf{z} \in M_1 \times M_2 \times .... \times M_n$. Afin de simplifier ce modèle, nous supposons par la suite que les n composantes catégorielles de $\mathbf{z} = (z^1, z^2, ..., z^k, ..., z^n)$ sont indépendantes, ce qui nous permet d'écrire $p(\mathbf{z}/c_1) = \prod_{k=1}^{n} p(\mathbf{z}^k/c_1)$ où $p(\mathbf{z}^k/c_1)$ est une table de probabilité de dimension m_k. Nous associons dans la suite à chaque cellule c_1 de la carte, n tables unidimensionnelles de probabilité. Chaque table de probabilité $p(z^k/c_1)$ contient les probabilités des m_k modalités de la composante z^k. Cette table de probabilité sera notée par la suite θ^{k,c_1} :

$$\theta^{k,c_1} = \{\theta_j^{k,c_1}, j = 1 \dots m_k\} \text{ avec } \theta_j^{k,c_1} = p(z^k = x_j^k / c_1).$$

L'ensemble des paramètres permettant de définir la table de probabilité d'une cellule c_1 de la carte $\mathcal{C}_1$, est constitué de l'union de toutes les tables de probabilités des variables composantes : $\theta^{c_1} = \cup_{k=1}^{n} \theta^{k,c_1}$.

Ainsi, l'ensemble des paramètres $\theta = \theta^{C_1} \cup \theta^{C_2}$ qui permettent de définir le modèle probabiliste de carte topologique est constitué par : l'ensemble des coefficients des tables : $\theta^{C_1} = \cup_{c=1}^{N_{cell}} \theta^{c_1}$, et l'ensemble des probabilités a priori : $\theta^{C_2} = \{\theta^{c_2}, c_2 = 1..N_{cell}\}$, où $\theta^{c_2} = p(c_2)$.

Le problème maintenant est de définir la fonction de coût et l'algorithme d'apprentissage qui permet d'estimer l'ensemble de ces paramètres.

2.2 Algorithme d'apprentissage CTM

On suppose que l'on dispose d'une base d'apprentissage $\mathcal{A} = \{\mathbf{z}_i; i = 1..N\}$ où toutes les observations $\mathbf{z}_i$ sont indépendantes. L'algorithme d'apprentissage consiste à maximiser la vraisemblance des observations en appliquant l'algorithme EM. L'usage de l'algorithme EM s'explique par l'existence d'une variable cachée notée ξ, constituée par le couple de cellule c_1 et c_2, $\xi = (c_1, c_2)$, responsable de la génération d'une donnée observée $\mathbf{z}$. En effet, la variable cachée $\xi = (c_1, c_2)$ apparaît lorsqu'on écrit :

$$p(\mathbf{z}) = \sum_{\xi \in \mathcal{C}_1 \times \mathcal{C}_2} p(\mathbf{z}, \xi) = \sum_{c_1 \in \mathcal{C}_1, c_2 \in \mathcal{C}_2} p(\mathbf{z}/c_1) p(c_1/c_2) p(c_2)$$

selon le formalisme probabiliste $p(\mathbf{z}, c_1, c_2) = p(c_2) p(c_1/c_2) p(\mathbf{z}/c_1)$

A chaque donnée réellement observée $\mathbf{z}$, il correspond une donnée catégorielle disjonctive non observée ξ qui appartient à $\mathcal{C}_1 \times \mathcal{C}_2$. Si l'on code cette variable par le codage binaire disjonctif, on obtient un vecteur binaire $\mathbf{y}$ de dimension $N_{cell} \times N_{cell}$ dont les composantes $y_{(c_1,c_2)}$ sont définies par :

$$y_{(c_1,c_2)} = \begin{cases} 1 & \text{si } \xi = (c_1, c_2) \\ 0 & \text{sinon} \end{cases}$$

Avec cette notation, $p(\mathbf{z}, \xi)$ s'écrit :

$$p(\mathbf{z}, \xi) = \prod_{c_2 \in \mathcal{C}_2} \prod_{c_1 \in \mathcal{C}_1} \left[p(c_2) \frac{K^T(\delta(c_2, c_1))}{T_{c_2}} p(\mathbf{z}/c_1) \right]^{y_{(c_1,c_2)}}$$

On note par y_i la représentation binaire disjonctive de la la variable cachée ξ_i relative à l'observation $\mathbf{z}_i$. On note par Ξ l'ensemble de ces variables cachées. La vraisemblance des observations avec les variables cachées s'écrit :

$$V^T(\mathcal{A}, \Xi; \theta) = \prod_{i=1}^{N} \prod_{c_2 \in \mathcal{C}_2} \prod_{c_1 \in \mathcal{C}_1} \left[\theta^{c_2} \frac{K^T(\delta(c_2, c_1))}{T_{c_2}} p(\mathbf{z}_i/c_1) \right]^{y_{i(c_1,c_2)}}$$

Le log-vraisemblance s'écrit :

$$\ln V^T(\mathcal{A}, \Xi; \theta) = \sum_{\mathbf{z}_i \in \mathcal{A}} \sum_{c_2 \in C_2} \sum_{c_1 \in C_1} y_{i_{(c_1,c_2)}} \left[\ln(\theta^{c_2})) + \ln(\frac{K^T(\delta(c_2, c_1))}{T_{c_2}}) + \ln(p(\mathbf{z}_i/c_1)) \right].$$

L'application de l'algorithme EM [4] pour la maximisation de la vraisemblance des données observées fournit un algorithme itératif où les paramètres à l'itération $t+1$ sont calculés en fonctions des paramètres estimés à l'itération t. Les formules calculant les paramètres à l'itération $t+1$ sont définies par :

$$p(c_2) = \theta^{c_2} = \frac{\sum_{\mathbf{z}_i \in \mathcal{A}} p(c_2/\mathbf{z}_i, \theta^t)}{N} \tag{1}$$

$$p(z^k = x_j^k/c_1) = \theta_j^{k,c_1} = \frac{\sum_{\mathbf{z}_i \in \tau_{k,j}} p(c_1/\mathbf{z}_i, \theta^t)}{\sum_{\mathbf{z}_i \in \mathcal{A}} p(c_1/\mathbf{z}_i, \theta^t)} \tag{2}$$

Avec

$$p(c_1/\mathbf{z}, \theta^t) = \frac{p(\mathbf{z}/c_1, \theta^t) \sum_{c_2 \in \mathcal{C}_2} p(c_2/\theta^t) p(c_1/c_2)}{p(\mathbf{z}/\theta^t)}, \tag{3}$$

$$p(c_2/\mathbf{z}, \theta^t) = \frac{\sum_{c_1 \in \mathcal{C}_1} p(c_2/\theta^t) p(c_1/c_2) p(\mathbf{z}/c_1, \theta^t)}{p(\mathbf{z}/\theta^t)} \tag{4}$$

et

$$\tau_{k,j} = \{\mathbf{z}_i \in \mathcal{A}; z_i^k = x_j^k\}$$

Celui-ci représente l'ensemble des observations $\mathbf{z}_i$ qui ont répondu par la modalité x_j^k à la k^{ieme} composante.

L'algorithme CTM [11] pour un paramètre T fixé se présente de la manière suivante :

Initialisation (itération $t = 0$)
détermination de l'ensemble des paramètres initiaux (θ^0) et du nombre d'itérations N_{iter}.

Itération de base à T constant (itération $t \geq 1$)
Calcul de l'ensemble des paramètres θ^{t+1} à partir de l'ensemble des paramètres précédent θ^t en appliquant les formules (2) et (1).

Répéter l'itération de base jusqu'à $t > N_{iter}$.

Dans ce paragraphe, nous avons présenté l'algorithme d'apprentissage CTM permettant d'estimer les paramètres maximisant la fonction log-vraisemblance pour un paramètre T fixé. Le paramètre T permet de contrôler la taille du voisinage d'influence d'une cellule sur la carte, celle-ci décroît avec le paramètre T. Par analogie avec l'algorithme des cartes topologiques de Kohonen, on peut faire décroître la valeur de T entre deux valeurs T_{max} et T_{min}. Pour chaque valeur de T, on obtient une fonction de vraisemblance V^T donc l'expression varie avec T. On peut observer deux étapes dans le fonctionnement de l'algorithme :

- La première étape correspond aux grandes valeurs de T. Dans ce cas, le voisinage d'influence d'une cellule c de la carte est grand et correspond aux valeurs "significatives" de $K^T(\delta(c,r))$. Les formules (1) et (2) ont tendance, à faire participer un très grand nombre d'observations à l'estimation des paramètres du modèle. Elle permet la mise en place de l'ordre topologique.

- La deuxième étape correspond aux petites valeurs de T. Le nombre d'observations intervenant dans les formules (1) et (2) est alors restreint. L'adaptation est alors très locale.

Si on utilise cette décroissance de T, l'algorithme d'apprentissage de CTM se présente de la manière suivante :

Phase d'initialisation (itération $t = 0$)
Choisir T_{max}, T_{min} et N_{iter}. Effectuer l'algorithme d'apprentissage CTM pour la valeur de T constante égale à T_{max}.

Etape itérative
L'ensemble des paramètres θ^t de l'étape précédente est connu. Calculer la nouvelle valeur de T en appliquant la formule suivante :

$$T = T_{max}\left(\frac{T_{min}}{T_{max}}\right)^{\frac{t}{N_{iter}-1}}$$

Pour cette valeur du paramètre T, appliquer l'itération de base décrite dans l'algorithme précédent, en appliquant les formules (1) et (2).

Répéter l'étape itérative tant que $t \leq N_{iter}$

2.3 CTM et Classification

Les calculs effectués lors du déroulement de l'algorithme CTM permettent d'éstimer l'ensemble des paramètres $\theta = \theta^{C_1} \cup \theta^{C_2}$ définis par le modèle probabiliste des cartes topologiques. Ces paramètres sont constitués par : l'ensemble des coefficients des tables : $\theta^{C_1} = \cup_{c=1}^{N_{cell}} \theta^{c_1}$, et l'ensemble des probabilités a priori : $\theta^{C_2} = \{\theta^{c_2}, c_2 = 1..N_{cell}\}$,(voir §2.1). Il permettent d'estimer la probabilité a posteriori $p(c_1/\mathbf{z}, \theta^t)$ par la formule suivante :

$$p(c_1/\mathbf{z},\theta) = \sum_{c_2 \in C_2} p(c_1, c_2/\mathbf{z},\theta) \tag{5}$$

avec

$$p(c_1/\mathbf{z},\theta) = \frac{p(\mathbf{z}/c_1,\theta)\sum_{c_2\in C_2}\theta^{c_2}K^T(\delta(c_1,c_2))}{\sum_{c_2\in C_2}\theta^{c_2}p_{c_2}(\mathbf{z})}$$

Cette probabilité a posteriori permet d'affecter une nouvelle observation $\mathbf{z}$ à la cellule de la carte ayant la probabilité a posteriori la plus grande. Il est à noter que l'introduction de cette étape dans l'algorithme d'apprentissage CTM permet d'obtenir une version classifiante [3].

Ces probabilités a posteriori permettent de faire de la classification. En effet, si les données sont étiquetées par un expert, à chaque observation $\mathbf{z}_i$ est associée l'étiquette l_i de sa classe, il est alors possible de calculer les probabilités a posteriori $p(l_i/\mathbf{z})$ permettant d'appliquer la règle de décision de Bayes. Si on note par n_{c_1} le nombre d'observations de $\mathcal{A}$ qui sont affectées à c_1 par la fonction d'affectation $\chi(\mathbf{z}) = \arg\max_{c_1} p(c_1/\mathbf{z})$ et par $n_{c_1}^{l_i}$ le nombre de ces observations étiquetées par l_i, on peut calculer la probabilité a posteriori de chaque classe avec la formule suivante :

$$p(l_i/\mathbf{z}) = \sum_{c_1\in\mathcal{C}_1} p(l_i/c_1)p(c_1/\mathbf{z}) \tag{6}$$

avec $p(l_i/c_1) = \frac{n_{c_1}^{l_i}}{n_{c_1}}$.

3 Exemple d'application

3.1 Exemple 1

Cet exemple dont les données proviennent d'une base d'assurance va permettre de montrer d'une part le bon fonctionnement de l'algorithme CTM et d'autre part son pouvoir de classification. La base utilisée contient 1106 individus. Chacun des individus est codé avec un vecteur de 9 composantes : **U**tilité (**Priv**é, Pro**f**essionel), **S**exe(**H**omme, **F**emme, **V**éhicule de **S**ociété), **Lang**ue (**Fr**ançais, **Au**tre), **Ag**e(**V**ieux, **M**oyen, **J**eune) **Lo**calisation (**C**apitale, **Pr**ovince), **B**onus(1,2), **Po**lice (86, **Au**tre), **Pu**issance(**Gr**ande, **Pet**ite), **A**ge **V**éhicule (**An**cien, **Nou**veau). Pour chaque individu, l'assurance a indiqué s'il s'agit d'un bon ou d'un mauvais conducteur.

Les premières expériences vont permettre d'illustrer le comportement de l'algorithme. Nous utiliserons dans un deuxième temps les étiquettes " bon conducteur" (1) et "mauvais conducteur" (2) pour montrer les capacités de classification de CTM. Nous testerons enfin ses capacités de généralisation.

L'apprentissage, d'une carte de dimension 5×5 cellules, effectué sur la base entière des 1106 individus, fournie pour chaque cellule, 9 tables de probabilités (§2.1) de dimension

deux ou trois ($\theta^{c_1} = \cup_{k=1}^{9} \theta^{k,c_1}$, $\theta^{k,c_1} = \{\theta_j^{k,c_1}, j = 1..2/3\}$).

Des visualisations simples sont réalisées à partir de la carte obtenue après apprentissage à l'aide de l'algorithme CTM. La figure 1 présente la table de probabilités estimées pour la première cellule située en haut et à gauche de la carte. On remarque que certaines modalités sont très probables. L'analyse des valeurs obtenues pour les probabilités nous permet d'interpréter cette cellule comme représentant les individus qui sont **Pr**ofessionnels (variable **U**) avec une probabilité de 0.99, qui vivent en **Pr**ovince (variable **Lo**) avec la probabilité de 0.85 et qui ont un **An**cien véhicule (variable **AV**) avec la probabilité 0.81. On constate que ces individus ont le premier bonus 1 (variable **B**) avec une probabilité de 0.98.

U	S	Lg	Ag	Lo	B	Po	Pu	AV
0.01	0.67	0.68	0.63	0.15	0.98	0.75	0.43	0.81
0.99	0.32	0.32	0.13	0.85	0.02	0.25	0.57	0.19
	0.01		0.24					

FIG. 1 – Les 9 tables de probabilités associées à la cellule située en haut à gauche de la carte

A la fin de la phase d'apprentissage, chaque observation représentant un assuré est affectée à la cellule ayant la plus forte probabilité a posteriori $p(c_1/\mathbf{z})$ (expression (5)). La figure 2 montre les 25 probabilités a postériori calculées sur toute la carte 5×5 pour une observation de la base $\mathbf{z} = (Pf, H, Fr, V, Pr, 1, 86, Pt, Nou)$. On constate sur la figure 2 que la distribution de probabilités $p(c_1/\mathbf{z})$ est une région connexe autour de la cellule la plus probable (couleur noire).

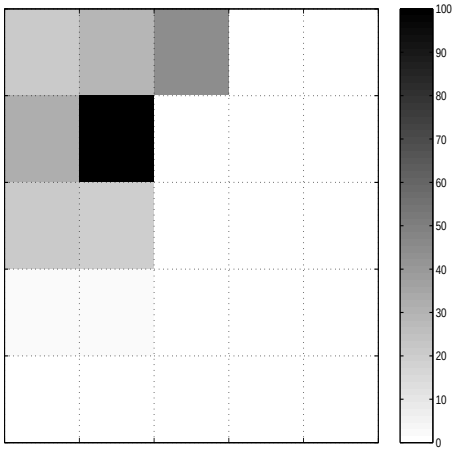

FIG. 2 – La probabilité a postériori $p(c_1/\mathbf{z})$ d'une observation $\mathbf{z}$

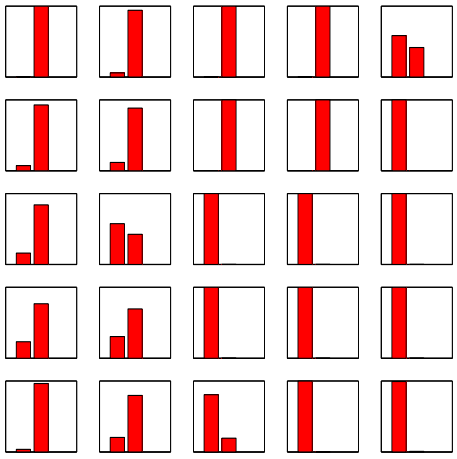

FIG. 3 – Distribution de la variable aléatoire "Utilité" du véhicule sur la carte 5×5, chaque cellule de la carte est repésentée par un histogramme ; la première barre indique la modalité **Privé** ; la deuxième barre indique la modalité **Professionnelle**

Une visualisation d'une variable peut être réalisée à l'aide des tables de probabilités. On peut visualiser la distribution de chacune des neuf variables aléatoires. La figure 3 montre la distribution des deux modalités de la variable aléatoire "Utilité" du véhicule sur la carte 5×5 sous forme d'histogramme. On observe une cohérence entre l'amplitude des 2 modalités et la structure topologique de la carte.

La figure 4 montre la distribution des trois modalités de la variable aléatoire **Age**. Le niveau de gris représente la probabilité $\theta^{Age,c_1} = p(z^{Age}/c_1)$ des trois modalités sur toute la carte. La carte 4.(V), correspondant à la probabilité θ_V^{Age,c_1}, indique que la partie gauche de la carte est dédiée aux conducteurs âgés avec une forte probabilité ; la carte 4.(M) de la probabilité θ_M^{Age,c_1} indique que les cellules en bas de la carte sont dédiées au conducteurs ayant un âge moyen. Les conducteurs jeunes sont représentés dans la carte 4.(J) par une région avec une forte probabilité (θ_J^{Age,c_1}).

Visualisation multi-dimensionnelle

Il est possible d'effectuer également une visualisation multi-dimensionnelle des résultats de CTM. On va pouvoir, en visualisant les variables, vérifier aussi le bon comportement de l'ordre topologique. Nous avons choisi de présenter les quatre tables de probabilités des variables qualitatives suivantes : **S**exe, **Age**, **Pu**issance, **A**ge **V**éhicule.

La figure 5 présente la distribution des quatre variables aléatoires sous forme d'histogramme. Pour chaque cellule de la carte on a représenté les quatre variables définies précédemment. Afin d'affiner notre analyse, nous avons représenté dans la figure 6 pour chaque cellule la modalité de la variable ayant la plus forte probabilité ($j_0 = \arg\max_j(\theta_j^{k,c_1})$). Cette probabilité ne figure que si elle est supérieure à 0.8

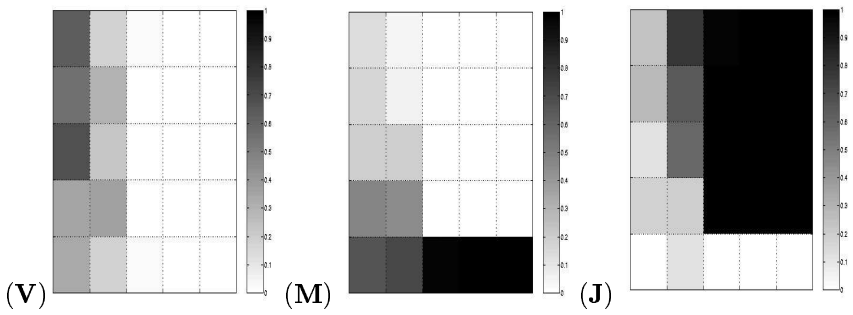

FIG. 4 – Carte topologique représentant la distribution des trois modalités de la variable aléatoire Age (**V** : **V**ieux, **M** : **M**oyen Age, **J** : **J**eune)

($\theta_{j_0}^{k,c_1} \geq 0.8$) dans le cas des variables à deux modalités et supérieure à 0.60 ($\theta_{j_0}^{k,c_1} \geq 0.6$) pour les variables avec trois modalités.

Les cellules positionnées en haut à droite de la carte représentent des jeunes conducteurs ayant majoritairement des véhicules neufs ; les conducteurs âgés sont localisés dans la partie gauche de la carte ; le reste des conducteurs ayant un âge moyen se retrouve en bas de la carte.

Afin de visualiser la cohérence de la carte avec l'étiquetage des observations (bon, mauvais), nous présentons à la figure 7 la carte obtenue après avoir effectuer un vote majoritaire sur les sous ensembles d'observations affectées aux cellules de la carte. Nous rappelons que dans ce cas la fonction d'affectation est définie par $\chi(\mathbf{z}) = \arg\max_{c_1} p(c_1/\mathbf{z})$ On distingue deux régions sur la carte qui sont dédiées aux deux types d'assurés. Les cellules en haut à gauche de la carte sont dédiées aux assurés n'ayant jamais d'accident (étiquetés par $''1''$) ; les cellules étiquetées par 2 sont dédiées aux assurés ayant commis au moins un accident. Les cellules sans étiquette présentent des cellules vides n'ayant capté aucune observation de l'ensemble d'apprentissage. Il est maintenant possible en regardant à la fois la figure 6 et 7 de voir que les bon conducteurs (qui n'ont jamais eu d'accident) sont majoritairement des jeunes avec des véhicules anciens. On peut voir aussi que les mauvais conducteurs ont fait des accidents avec des véhicules puissants. Les mauvais conducteurs sont constitués majoritairement par des personnes jeunes et des personnes ayant un âge moyen.

Afin de tester les performances de l'algorithme d'apprentissage CTM en classification, nous avons repris la base complète des assurés constituée de 1106 individus. Nous avons découpé la base en trois sous-ensembles de même taille notés B_1, B_2 et B_3. Ainsi, nous avons effectué trois apprentissages différents en utilisant deux ensembles à la fois. La première carte est apprise avec les bases B_1 et B_2, la deuxième avec les deux bases B_1 et B_3 et finalement la troisième avec B_2 et B_3. Afin de tester les performances de chacune des 3 cartes comme classifieur, nous avons classé à l'aide de

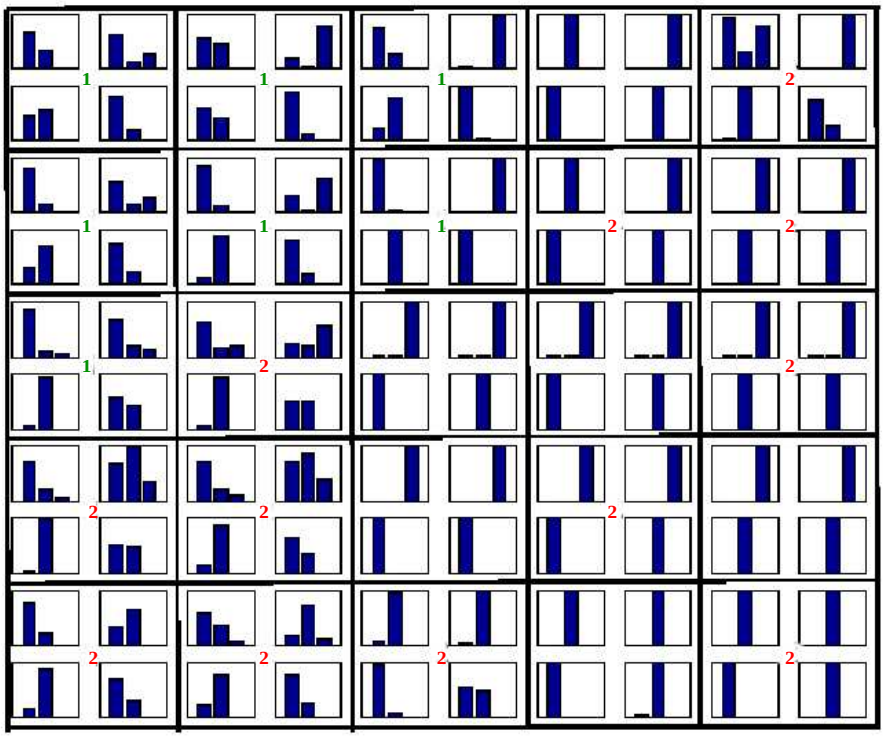

FIG. 5 – Distribution de la probabilité de quatre variables aléatoires ; chaque cellule de la carte est représentée par 4 histogrammes ; dans chaque cellule, la ligne du haut à gauche présente la variable “Sexe” qui correspond au premier histogramme, sur la même ligne, on a la variable “Age” ; sur la deuxième ligne, on a la variable “Puissance” à gauche suivie de la variable “Age Véhicule”

H V - An	- J - An	H J - An	F J Pt Nou	- J Gr-
H - - -	- J - An	H J Gr An	F J Pt Nou	VS J Gr Nou
H V Gr -	H - Gr -	VS J Pt Nou	VS J Pt Nou	VS J Gr Nou
- - Gr -	H - Gr -	VS J Pt An	VS J Pt An	VS J Gr Nou
H M Gr -	- M - -	F M Pt -	F M Pt Nou	F M Pt Nou

FIG. 6 – Carte 5×5, pour chaque cellule on affiche les quatres modalités ayant la plus forte probabilité. H : Homme, F : Femme, J : Jeune, M : âge Moyen, V : Vieux, VS : Véhicule de Service, An : Ancien véhicule, Nou : Nouveau véhicule. Gr : Grande puissance, Pt : Petite puissance.

1	1	1	-	2
1	1	1	2	2
1	2	-	-	2
2	2	-	2	-
2	2	2	-	2

FIG. 7 – Carte étiquetée après application du vote majoritaire, les cellules sans étiquette représentent des sous-ensembles vides. 1 : bon conducteur, 2 : mauvais conducteur

chaque carte la base qui n'a pas servi à l'apprentissage de celle-ci. Le tableau 1 fournit le taux de bonne classification avec les trois cartes.

Carte 1	Carte 2	Carte 3	Moyenne
86%-	84% -	85%	85%

TAB. 1 – Taux de bonne classification. Carte 1 : base d'apprentissage : B_1 et B_2, base de test : B_3 ; Carte 2 : base d'apprentissage : B_1 et B_3, base de test : B_2 ; Carte 3 : base d'apprentissage : B_2 et B_3, base de test : B_1

3.2 Exemple 2 : Construction du Corpus de Mots représentatifs

Le deuxième exemple analyse une base extraite d'une enquête de sémiologie portant sur le sens des mots chez différentes personnes. Cette enquête est constituée par un ensemble de mots possédant un certain nombre de qualités définies par l'auteur [12, 16] : Univocité sémantique (chaque mot ne doit posséder qu'un seul sens premier), Stabilité sémantique (le sens de chaque mot est stable dans le temps, c'est-à-dire qu'un mot ne peut pas posséder un sens pour un groupe et un autre pour les autres), Non-consensualité (les mots doivent être capables de provoquer des réactions contradictoires) et intensité émotionnelle (les mots doivent posséder une charge affective suffisante afin de pouvoir jouer le rôle de stimuli).

L'enquête consiste à demander aux individus de noter les mots en fonction du plaisir que l'ensemble de leurs connotations fait naître. Le but de l'enquête est de définir des groupes d'individus homogènes par rapport aux réponses apportées à l'enquête et de déterminer par la suite à l'aide de variables additionelles, les caractéristiques de ces groupes.

Ces données nous ont été fournies par Monsieur L. Lebart ; cette enquête est issue d'interrogations effectuées par la SOFRES auprès de son panel télématique. Nous avions accès à 1128 individus représentatifs de la population française. Cette enquête portait sur 286 mots, dans le cadre de cette étude nous n'avons eu accès qu'à 70 mots. Chaque individu interrogé devait donner une note comprise entre 1 et 7 permettant d'évaluer sa réaction à ce mot.

Afin de vérifier la cohérence locale de l'espace, l'auteur de la base introduit le chanps sémantique d'un mot en calculant les n mots les plus proches au sens d'une distance dérivée du coéfficient de corrélation. Pour le mot *Mystère*, le chanps sémantique est constitué par la liste suivante : *Inconnu, Orage, Secret, Sauvage, Aventurier, Emotion, Original, Magie, Nuit et Changement*, (ces mots sont écrits dans l'ordre croissant des distances).

L'algorithme CTM permet d'effectuer l'analyse recherchée : les notes affectées aux mots seront donc considérées comme des variables catégorielles. Pour CTM, à chaque cellule

est associée 70 tables unidimentionnelles de dimension 7. La k^{ieme} table représente les probabilités de la note affectée au k^{ieme} mot : on a donc pour chaque cellule c_1 et pour chaque k^{ieme} mot, la table de probabilités $\theta^{k,c_1} = \{\theta_j^{k,c_1}, j = 1..7\}$ (voir §2). L'ensemble de tous les paramètres estimés pour chaque cellule c_1 est égal à $\theta^{c_1} = \cup_{k=1}^{70} \theta^{k,c_1}$ auxquel il faut ajouter les coefficients du mélange correspondant aux cellule c_2 de la "deuxième" carte $\mathcal{C}_2$ noté θ^{c_2}.

L'apprentissage d'une carte CTM de 7×7 cellules est effectuée, et il permet d'estimer les paramètres des 70 mots dont nous disposons. L'apprentissage a été fait dans les conditions suivantes : à l'initialisation $\theta^{c_2} = \frac{1}{N_{cell}}$ et $\theta_j^{k,c_1} = \frac{\text{nombre d'élément de } \tau_{k,j}}{N_{cell}}$. $T_{max} = 1$ et $T_{min} = 0.1$ avec $N_{iter} = 1000$. A la fin de l'apprentissage si on affecte chaque observation $\mathbf{z}_i$ à la cellule $c = arg\max_{c_2 \in C_2} p(c_2/\mathbf{z}_i)$ on obtient la répartition des observations présentée par la figure 8. On observe que la partition obtenue a permis de bien distribuer les observations sur les 49 cellules de la carte.

38	30	39	43	28	26	15
30	20	26	39	16	29	12
26	28	26	31	21	18	16
35	10	18	26	13	65	18
29	26	23	19	46	62	44
20	14	21	15	18	51	21
3	15	13	11	9	10	16

FIG. 8 – Cardinalité des sous-ensembles, carte CTM 7×7

A l'aide de cette carte 7×7 et des informations probabilistes qui lui sont associées, il est possible d'effectuer un certain nombre d'analyses de la base étudiée. Etant donné que le but des expériences qui suivent est de montrer le bon fonctionnement de CTM, nous nous sommes limités à étudier les effets dûs à quelques mots pour lesquels l'exactitude des propriétés sémantiques retrouvées peuvent être vérifiées.

La figure 9 illustre les distributions de probabilités des notes attribuées au mot *mort*, sous forme d'une carte en niveaux de gris. Chaque carte représente la distribution d'une note j ($j = 1..7$) correspondant à la probabilité estimée $\theta_j^{Mort,c_1} = p(z^{Mort} = j/c_1)$. On remarque que la probabilité est très forte dans la carte correspondant à la note "1" (la carte N1 de la figure 9) ; on constate que l'ensemble des individus pris dans la base n'apprécient pas la *Mort*. Un nombre restreint d'individus accepte facilement la *Mort*, ils apparaissent sur les cartes 9.(N2), 9.(N3) et 9.(N4), mais avec une faible probabilité qui n'excède pas 0.3 ($\theta_2^{Mort,c_1} \leq 0.3, \theta_3^{Mort,c_1} \leq 0.3, \theta_4^{Mort,c_1} \leq 0.3$). Les notes 5, 6 et 7

sont présentées par des probabilités très faibles. Le mot *Mort* présente globalement, une stabilité sémantique. Une même analyse peut être effectuée sur chacun des 70 mots. La carte résultat de l'algorithme d'apprentissage CTM permet d'obtenir une visualisation (et une information numérique) sur les distributions des différentes notes attachées à un mot donné.

La figure 10 représente la distribution de probabilités de la note 1 pour le mot *Guerre*. Cette carte montre une mauvaise appréciation de ce mot par la majorité des individus de la base. Le mot *Guerre* se distingue par une stabilité sémantique. La carte N1 de la figure 9 représente la distribution de probabilités de la note 1 du mot *Mort*, on constate une grande similarité entre les représentations pour le mot *Guerre* et *Mort*, ce qui peut laisser penser que l'ensemble des individus qui compose l'enquête réagit d'une manière similaire aux deux mots. Les groupes de personnes qui n'apprécient pas la *Mort*, n'apprécient pas non plus la *Guerre* (une cellule avec une plus forte probabilité dans la carte N1 de la figure 9 se retrouve avec une plus forte probabilité dans la figure 10). On peut visualiser d'une seconde manière les probabilités estimées par CTM en affichant pour chaque mot uniquement la note qui apparaît avec la probabilité maximale. La figure 11 présente pour le mot *Immobile* et en chaque cellule de la carte, la classe j la plus probable et la probabilité associée. L'étiquette affectée à chaque cellule de la carte représente la note j correspondant à la probabilité maximale, sur la figure, le niveau de gris représente la probabilité maximale. On constate une homogénéité spatiale qui apparait au niveau des notes attribuées à ce mot malgré les probabilités très variées. La probabilité est plus forte pour toutes les cellules dont la la note est 4, cependant plus la zone sur la carte est "claire" et plus d'autres notes sont possibles.

La même étude est reprise pour le mot *Fleur* dans la figure 12.(a), cette carte présente pour chaque cellule et pour le mot *Fleur* la plus forte probabilité rencontrée dans la table de probabilités et la note qui lui est associée. On observe que la majorité des individus sont répartis entre ceux qui ont donné la note 6 et 7 à ce mot avec une forte probabilité pour la note 7. On constate qu'il existe une unicité sémantique du mot *Fleur*. La carte 12.(b) donne une même représentation du mot *Séduire*; on voit que la répartition de probabilités et de notes sont proches de celles du mot *Fleur*. Les individus qui aiment offrir des fleurs, cellules avec des notes élevées, représentés par une plus forte probabilité (cellules noircies), aiment bien séduire et ils sont représentés aussi par des cellules ayant des notes élevées avec une plus forte probabilité (figure 12.(a) et 12.(b)).

Nous pouvons aussi définir d'une manière plus globale la partition trouvée par CTM en définissant chaque sous groupe de cellules par un sous ensemble de mots qui le représente le mieux. Pour chaque cellule, nous affichons tous les mots qui contiennent dans leur table de probabilités une valeur supérieure à 0.8 ($k = \arg\{\theta_j^{k,c_1} \geq 0.8, j = 1..7\}$). Chaque cellule de la carte, présentée par la figure 13, est donc caractérisée par un ensemble de mots qui constituent un ensemble de sens et de qualités représentatifs du sous groupe d'individus affectés à la cellule. On remarque que les deux côtés (gauche et droite) de la carte opposent des groupes d'individus qui ont des réactions plus tranchées

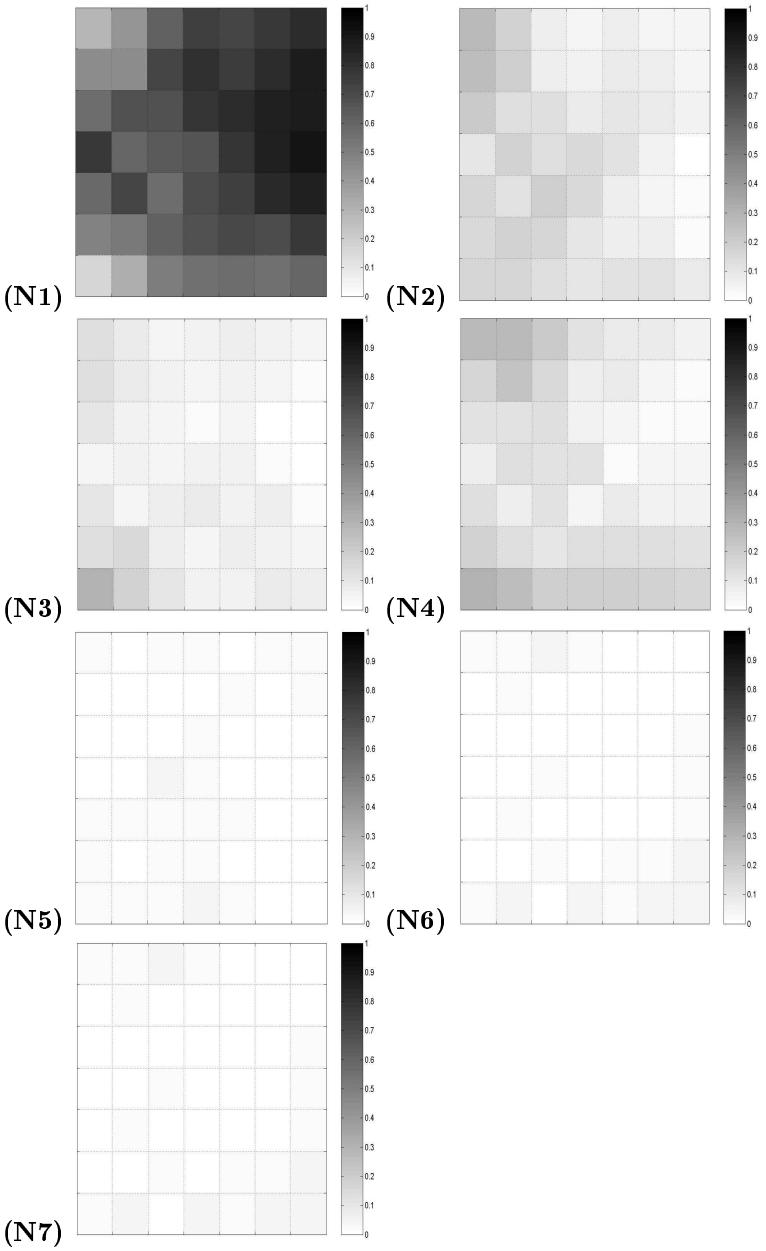

FIG. 9 – Cartes Topologiques décrivant les distributions de probabilités sur le mot Mort ; La carte (Nj) représente la distribution de probabilités correspondante à la note j

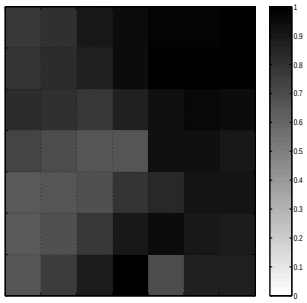

FIG. 10 – Carte Topologique donnant pour chaque cellule la probabilité estimée pour la note 1 du mot *Guerre*

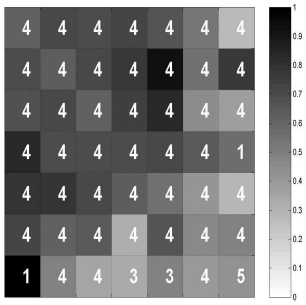

FIG. 11 – Carte Topologique donnat la note de la plus forte probabilité ainsi que la note associée pour le mot *Immobile*

(beaucoup de probabilités supérieures à 0.8) à d'autres groupes dont les avis sont plus mélangés.

Avec ce procédé, plusieurs visualisations peuvent être effectuées. Chaque seuil de probabilité représente un niveau d'investigation ; il est possible de faire apparaître les mots selon leurs caractéristiques sémantiques. En diminuant le seuil de probabilité de 0.8 à 0.6 et en ne tenant compte que des mots qui ont la même note "1", on obtient une nouvelle carte présentée dans la figure 14. On observe que les mots *Guerre* et *Mort* sont majoritairement mal appréciés par tous les individus tels qu'ils étaient dans la carte 13. On observe de plus le mot *Angoisse* qui apparaît sur la carte 14, d'autres mots tels que *Soldat*, *Punir* et *Immobile* apparaissent en diminuant le seuil de probabilités. Ceux-ci sont du même champ sémantique que *Guerre*.

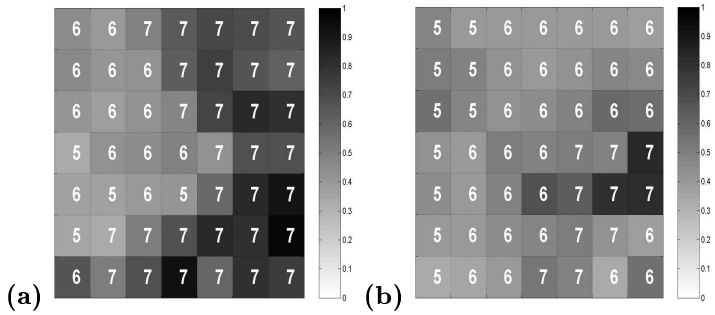

FIG. 12 – Comparaison des plus fortes probabilités et de notes associées pour le mot *Fleur* (a), et le mot *séduire* (b)

4 Conclusion

Dans ce papier, nous avons présenté un algorithme de carte topologique dédiée aux données catégorielles. Cet algorithme CTM se base sur le formalisme probabiliste des cartes topologiques et utilise l'algorithme EM pour maximiser la vraisemblance. Les expériences présentées montrent la robustesse de celles-ci à traiter des données dont les variables peuvent avoir plus de deux modalités. D'autre part, nous avons vu qu'à travers les divers aspects de visualisations, l'algorithme CTM fourni une quantités d'informations exploitables dans des applications réelles. Toutes les visualisations qui ont été effectuées montrent que l'ordre topologique permet d'obtenir une interprétation des groupements obtenus.

Références

[1] Anouar, F. Badran, F. Thiria, S. (1998) : Probabilistic self-organizing map and radial basis function networks. Neurocomputing 20, 83-96.

[2] Bishop, C. M., Svensen, M., and Williams, C. K. I. (1998) : GTM : The Generative Topographic Mapping. Neural Computation, 10(1), 215-234.

[3] Celeux, G. Govaert, G. A classification EM algorithm for clustering and stochastic version. Computational Statistics and Data analysis. Vol. 14, pp.351-332, 1992.

[4] Dempster,A. P. Laird, N. M. Rubin, D. B. Maximum liklihoode from incomplete data via the EM algorithm. Journal of royal Statistic Society, Series B, 39, 1-38

[5] Dolinicar, Weingessel , A. Buchta, C. (1998) : Dimitriadou, E. A Comparaison of several cluster algorithms on artificial binary data, scenarios from travel market segmentation. Working paper series 19, SFB (adaptive information systems and modelling in economics and management science).

[6] Kaban, A and Girolami, M. (2001) : A Combined Latent Class and Trait Model for the Analysis and Visualisation of Discrete Data. I.E.E.E Transactions on Pattern Analysis and Machine Intelligence.23(8), pp859 -.872.

	Guerre	Guerre	Guerre	Honnête, Guerre	Honnête, Guerre	Honnête, Guerre
	Guerre	Guerre	Guerre	Etranger Souverain Immobile Guerre	Guerre Honnête	Cadeau, Courage Guerre, Honnête Infini, Moëlleux Mort, Politesse
Guerre	Guerre		Guerre Mort	Guerre Immobile	Fleur Guerre Maison	Cadeau, Courage Dynamique, Fleur Guerre, Honnête Mort, Respect Victoire
Immobile				Angoisse Guerre Politesse	Cadeau Guerre Mort	Humour, Cadeau Courage, Gloire Séduire Maison, Mort Océan, Parfum Respect, Richesse Guerre, Angoisse
				Guerre	Cadeau Courage Fleur Séduire Honnête Maison Mort Guerre	Argent, Bijou Cadeau, Courage Dynamique, Fleur Guerre, Honnête Humour, intime Maison, Mort Politesse, Protéger Respect, Richesse Séduire, Victoire
			Guerre	Fleur Guerre	Fleur Guerre	Charitable, Courage Discipline, Dynamique Fleur, Guerre Honnête, Loi Maison, Politesse Protéger, Respect Travail,Victoire
Dynamique, Humour Peau, Intime Immobile		Guerre, Humour	Fleur, Humour Intime, Guerre		Fleur Guerre	Gloire Guerre

FIG. 13 – Carte Topologique 7×7. Chaque cellule contient l'ensemble des mots pour lesquels la probabilité maximale est supérieure à 0.8

Guerre	Guerre	Guerre	Guerre, Mort	Angoisse, Guerre Mort	Angoisse, Guerre Mort	Angoisse, Guerre Mort
Guerre	Guerre	Guerre, Mort	Angoisse, Guerre Mort	Angoisse, Guerre Mort	Angoisse, Guerre Mort	Angoisse, Guerre Mort
Guerre	Guerre, Mort	Guerre, Mort	Guerre, Mort	Angoisse, Guerre Mort	Angoisse, Guerre Mort	Angoisse, Guerre Mort
Guerre, Mort	Guerre, Mort	Guerre, Mort	Guerre	Angoisse, Guerre Mort	Angoisse, Guerre Mort	Angoisse, Critiquer Guerre, Mort Punir
Guerre	Guerre, Mort	Guerre	Guerre	Guerre, Mort	Angoisse, Guerre Mort	Angoisse, Critiquer Mort
Guerre	Guerre	Guerre, Mort	Guerre, Mort	Angoisse, Guerre Mort	Guerre	Guerre, Mort
Angoisse, Guerre Immobile, Soldat	Guerre	Angoisse, Guerre Soldat	Angoisse, Guerre Soldat	Guerre	Guerre	Guerre

FIG. 14 – Carte Topologique 7×7. Chaque cellule contient l'ensemble des mots pour lesquels la note 1 obtient une probabilité supérieure à 0.6

[7] Ibbou, S. Cottrell, M. Multiple correspondance Analysis crosstabulation matrix using the Kohonen algorithm. In verlaeysen, M. Editor proc of ESANN'95, pages 27-32. Dfacto Bruxelles 1995.

[8] Kaski. S, Honkela. T, Lagus. K, and Kohonen. T. (1998). WEBSOM–self-organizing maps of document collections. Neurocomputing, volume 21, pages 101-117.

[9] Kohonen, T. (1994) : Self-Organizing Map. Springer, Berlin.

[10] Lebbah. M, Thiria. S, Badran. ESANN, Topological Map for Binary Data, ESANN 2000, Bruges, April 26-27-28, 2000, Proceedings.

[11] Lebbah. M , Thiria. S, Badran. F, Chabanon. C. ICANN 2002, Categorical Topological map, Madrid 2002.

[12] Lebart, L. Piron, M. Steiner J.-F. La sémiométrie. Dunod, Paris. 2003.

[13] Leich, F. Weingessel, A. Dimitriadou, E. (1998) : Competitive Learning for Binary Data. Proc of ICANN'98, septembre 2-4. Springer Verlag.

[14] Luttrel S. P. (1994). A Bayesian Ananlysis of Self-Organizing Maps, Neural Computing vol 6

[15] McLachlan, G. Krishman, T. The EM algorithm and Extensions. Wiley, New York, 1997.

[16] Steiner J.-F., Auliard, O. La sémiometrie : un outil de validation des réponses, In : La Qualité de l'Information dans les Enquêtes / Quality of Information in Sample Surveys, ASU, (L. Lebart ed.), Dunod, Paris, p 241-274, 1992.

Summary

This paper introduces a topological map dedicated to cluster analysis and visualization of categorical data. Usually, when dealing with categorical data, topological maps use an encoding stage : categorical data are changed into numerical vectors and traditional numerical algorithms are run. In the present paper, we propose a probabilistic formalism where cells are now represented by probability tables. Two examples using actual data allow to validate the approach. The results show the good quality of the topological order obtained as well as its performances in classification.