Visualisation des données textuelles

Rodolphe Priam[1], Annie Morin[1]

[1] IRISA – Campus de Beaulieu – 35042 Rennes cedex – France

Résumé

Le but de cet article est d'élaboration d'une méthode de projection en carte bidimensionnelle des données textuelles. Nous mettons l'accent sur la modélisation cohérente et efficace de ces données considérées comme des échantillons générés par une distribution multinomiale à déterminer. Nous adaptons un réseau de neurones particulier qui généralise à la fois l'analyse en composantes pricipales et la méthode des moyennes mobiles. De ce fait, nous aboutissons à un tout nouveau modèle probabiliste d'analyse des données qualitatives dont nous éclairons le fonctionnement en mettant en évidence l'erreur implicite minimisée. Nous décrivons également des moyens de représentation de la carte obtenue dont l'un d'eux est original, ainsi que des indicateurs comparables à ceux employés en analyse factorielle. Diverses expériences sur des données réelles permettent de constater la pertinence de l'approche. En conclusion, nous commentons les résultats obtenus et envisageons des travaux futurs.

Abstract

This work introduces a new way to visualize textual data by modelling texts as a mixture of multinomial distribution, and explains a model which generalizes PCA and K-means methods. The introduction presents briefly principal surfaces. Then, we describe the new model to map textual data and show how it works by calculating the implicit distance it uses. A review of alternative methods follows. We describe ways to project data with the model. After the proposal of indicators, several experiences over several corpus are presented. The conclusion summarizes the results obtained and gives some perspectives.

Mots-clés : ACP non-linéaire, variante des moyennes mobiles, carte auto-organisatrice de Kohonen, visualisation, classification non supervisée, distribution discréte, EM

1. Introduction

Nous nous intéressons aux méthodes de projection planaire telles que les projections factorielles. Ces dernières sont très utilisées en analyse des données textuelles pour leurs bons résultats. Elles sont aussi connues pour leur complexité algorithmique importante liée à la diagonalisation d'une matrice creuse, ainsi que pour leur difficulté d'interprétation en grandes dimensions, causée par les nombreuses cartes à visualiser ainsi que le bruit manifesté par des mots parfois mal représentés ou peu contributifs. Tout cela provient du caractère linéaire de la projection qui n'approche pas localement les données mais plutôt globalement. Nous étudions des méthodes d'approximation de ces projections de manière à les rendre non-linéaires tout en étant adaptées au traitement du langage naturel. Nous présentons dans la partie suivante un algorithme de projection efficace pour modèliser les différentes thématiques présentes dans un corpus ainsi que leur liaison statistique quantifiée par des relations de proximité sur un treillis à deux dimensions ou carte topologique. Auparavant, cette introduction présente succintement les surfaces principales, constructions sur lesquelles repose notre algorithme, ainsi que les notations nécessaires dans la suite.

Les surfaces principales (non-paramètriques) de (Hastie and Stuetzle, 1989) sont des ACPs non-linéaires ; elles modélisent une surface non-linéaire sur laquelle se projette la distribution des données, en généralisant la propriété de moindres-carrés de l'ACP. Les cartes auto-organisatrices de Kohonen sont développées par son auteur dans (Teuvo Kohonen, 1997) ; c'est l'algorithme du SOM pour *Self-Organizing Map*. L'algorithme du SOM est un exemple d'apprentissage d'une surface principale par discrétisation. Il se se définit comme un ensemble de neurones contraints en grille dessinant une surface discrète déformée dans l'espace des données. Après apprentissage, la grille ayant convergée vers une position stable, chaque neurone est par définition la moyenne (approchée) des individus qui lui sont le plus proche[1]. C'est pourquoi, le SOM généralise également la méthode des moyennes mobiles.

Il est commode de représenter le treillis de neurones en deux ou trois dimensions pour visualiser la projection obtenue. Ces représentations synthétiques offrent un moyen rapide et aisé d'accès à une information massive contrairement à l'ACP classique. Il existe (Bishop et al., 1998) quelques versions paramétriques du SOM, qui sont généralement adaptées à des données continues. Kohonen et son équipe ont développé une implémentation du SOM capable de projeter un nombre important de textes sur une carte. Les heuristiques employées dans cette projection rendent difficile l'interprétation des résultats obtenus. Nous nous intéressons à un moyen de projection offrant une visualisation comparable au SOM mais par contre dotée d'indicateurs statistiques permettant une meilleure interprétation des projections. Or, le modèle multinomial de document permet de construire un classifieur performant[2] appellé Naive Bayes en théorie de l'apprentissage. Nous étudions une projection non supervisée basée sur un mélange de multinomiales contraintes en carte topologique.

Un corpus de I documents est noté $\mathcal{D}$, un document est représenté par un vecteur d_i portant les fréquences N_{ij} d'occurrences des J mots m_j du vocabulaire $\mathcal{V}$; un document d_i est une suite[3] de mots "$m_{i1}, m_{i2}, \cdots, m_{i|d_i|}$" auquel on associe le vecteur de comptage $[N_{i1}N_{i2}\cdots N_{iJ}]$. Les classes de textes sont notées $x_k \in \mathcal{X}$, où x_k s'identifiera aussi bien à la variable aléatoire de label de classe, qu'à son centre dans l'espace des données, ou même à son vecteur de coordonnées sur la grille, suivant le contexte. Enfin, $P(.|\theta)$ représente une distribution sur l'espace des mots ou des documents, paramétrée par le vecteur θ. Au maximum de vraisemblance, on note $\theta = \hat{\theta}$. On pose $P^{\theta}_{\bullet|k}$ la distribution de $\mathcal{V}$ conditionnellement à x_k, elle correspond aux $P(m_j|x_k, \theta)$. On note les probabilités empiriques $P_{j|i} = \frac{N_{ij}}{N_{i\bullet}}$, $P_i = \frac{N_{i\bullet}}{N_{\bullet\bullet}}$. On pose $P_{\bullet|i}$, la distribution empirique de $\mathcal{V}$ conditionnellement à d_i.

2. Cartographie par multinomiales contraintes

Cette partie introduit un algorithme de cartographie associative à base de gaussiennes du à (Ambroise and Govaert, 1996) puis la méthode que nous proposons ainsi que les liens avec les moyennes mobiles. On appelle h, la fonction de voisinage qui prend en compte la topologie[4]

1. Il faut bien faire la différence entre la distance des neurones contraints sur la grille à deux dimensions et la distance réelle des neurones en tant que vecteur plongé dans un espace multidimensionnel identique à celui des vecteurs d'entrée du réseau.

2. Ce classifieur est compétitif devant des méthodes plus élaborées telles que les arbres de décision.

3. On ne conserve pas la ponctuation.

4. i.e. $h(x_{k_1}, x_{k^*}) \leq h(x_{k_2}, x_{k^*}) si ||x_{k_1} - x_{k^*}||_{\mathcal{R}^2} > ||x_{k_2} - x_{k^*}||_{\mathcal{R}^2}$ sur la grille.

de la grille. Par exemple $h(x_{k1}, x_{k2}) \propto exp(-\frac{1}{2\sigma^2}||x_{k1} - x_{k2}||^2_{\mathcal{R}^2})$. On note[5] un paramètre avec une étoile tout paramètre de la classe à laquelle d_i est affecté.

2.1. EM topologique et SOM

Dans cette section, d_i est un vecteur à valeurs continues. Nous résumons l'algorithme du *Topology Preserving EM* ou TPEM pour une cartographie associative. On se réferrera à l'article originel pour le développement original. Le TPEM est intéressant à étudier puisqu'il s'agit d'une variante probabiliste du SOM. Il permet l'estimation et l'organisation en carte de distributions quelconques même si les auteurs justifient la méthode seulement pour un mélange de gaussiennes $\mathcal{N}(x_k, \Sigma_k)$ de moyennes x_k et matrices de variance-covariance Σ_k. A notre connaissance, ils ont étudié seulement ce cas particulier. La méthode optimise la vraisemblance classifiante :

$$\mathcal{L}_{\mathcal{G}}(\mathcal{D}|\theta) = \sum_k \sum_i \mu_{ik} \ln[P(x_k|\theta)\mathcal{N}(x_k, \Sigma_k)] \text{ où } \mu_{ik} \in \{0, 1\} \quad (1)$$

Dans le cas de gaussiennes isotropiques équiprobables, l'estimation du mélange classifieur et la minimisation du critère moyennes mobiles (ou critère de variance intra-classe) sont équivalentes (cf.(Droesbeke et al., 1992)).

Les auteurs font alors le lien entre les moyennes mobiles et le SOM car sous les bonnes hypothèses de (Ambroise and Govaert, 1996), le SOM minimise[6] $W = \sum_k \sum_i h(x_k, x^*_k)||d_i - x_k||^2_{\mathcal{R}^J}$. Finalement, le TPEM est une variante d'EM classifiant (ou CEM, (Droesbeke et al., 1992)) dont la phase d'affectation est rendue[7] floue. La convergence est assurée car les derniers pas se transforment en algorithme CEM puisque les coefficients d'affectation μ_{ik} deviennent binaires.

2.2. Méthode proposée et comparaison avec les moyennes mobiles

Nous modifions l'algorithme précédent en remplaçant la distribution gaussienne par une multinomiale. Nous obtenons :

$$\mathcal{L}_{\mathcal{M}}(\mathcal{D}|\theta) = \sum_k \sum_i \mu_{ik} \sum_j N_{ij} \ \ln[P(m_j|x_k, \theta)] + \sum_k \sum_i \mu_{ik} \ \ln[P(x_k|\theta)] \quad (2)$$

Les justifications dans le cas gaussien ne sont plus valables puisque le critère $E_{||\cdot||}$ des moyennes mobiles n'est plus valide pour des variables discrètes. En supposant les classes équiprobables :

$$\mathcal{L}_{\mathcal{M}}(\mathcal{D}|\theta) = -N_{\bullet\bullet} \sum_k \sum_i \ P_i \ \mu_{ik} \ \{\sum_j P_{j|i} \ \ln[\frac{P_{j|i}}{P(m_j|x_k, \theta)}]\} + Cte \quad (3)$$

D'où l'équivalence, :

$$max_\theta \ \mathcal{L}_{\mathcal{M}}(\mathcal{D}|\theta) \Leftrightarrow min_\theta E_{KL} = \sum_i P_i \ \sum_k \ \mu_{ik} \ KL(P_{\bullet|i}||P^\theta_{\bullet|k}) \quad (4)$$

5. si il n'y a pas d'ambiguité possible

6. L'algorithme du SOM se termine en critère moyennes mobiles lorsque $h(x_k, x^*_k) = \delta(k, k^*)$; ce qui arrive au moment où la fonction h annule tout voisinage, au bout d'un certain nombre de pas d'apprentissage. En effet, en pratique, h n'est pas fixée et son rayon d'action au voisinage de chacune des classes est diminué sur la grille.

7. On remplace la phase d'affectation "$\mu_{ik^*} = 1$ ssi $P(x_{k^*}|\theta)\mathcal{N}(x_{k^*}, \Sigma_{k^*}) > P(x_k|\theta)\mathcal{N}(x_k, \Sigma_k), \forall k$ et $\mu_{ik} = 0$ sinon " par "$\mu_{ik} = \frac{h(x_k, x_{k^*})}{\sum_{\acute{k}} h(x_{\acute{k}}, x_{k^*})}$", sachant que $h(x_k, x_{k^*}) \to \delta(k, k^*) \Rightarrow \mu_{ik} \to \delta(k, k^*)$ en décroissant.

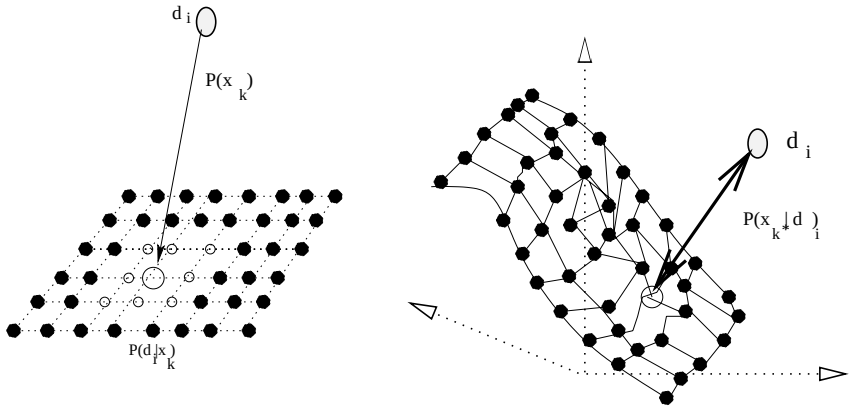

FIG. 1 – *Schéma formel et illustration d'une carte apprise. Sur la gauche de la figure, un schéma formel du réseau de neurones étudié est représenté. La sortie est la grille à deux dimensions. Le neurone en forme d'ellipse est l'entrée. Les traits en pointillés montrent les relations directes de voisinage. Sur la droite, un exemple de réseau plongé dans l'espace des données (en réalité, l'espace probabiliste associé, vue en 3 dimensions seulement) est illustré. La probabilité a posteriori de d_i est illustrée par la double flèche*

Le maximum de vraisemblance ne minimise plus directement une variance intra en terme de distance euclidienne entre des centres et un ensemble d'individus à classer. Il minimise la somme des divergences de Kullback-Leibler[8] (KL) entre la distribution empirique des textes et la distribution à calculer, pondérées par la probabilité empirique d'apparition de chacun des textes. L'expression[9] obtenue est très comparable au critère moyennes mobiles puisqu'il suffit de remplacer la distance euclidienne par la divergence, et d'ajouter le facteur de pondération. Le facteur de pondération P_i prend en compte la taille des textes, propriété à noter puisqu'elle rappelle la distance du χ^2.

En remplaçant la divergence par la distance du χ^2, on obtient un autre critère très proche de celui étudié, cette distance convient pour la comparaison des vecteurs de données qualitatives sans supposer des hypothèses de distribution multinomiale :

$$min_\theta E_{\chi^2} = \sum_i P_i \sum_k \mu_{ik} \sum_j \frac{1}{Pj}(P_{j|i} - P(m_j|x_k,\theta))^2 = \sum_i P_i \sum_k \mu_{ik}\, d_{\chi^2}(P_{\bullet|i}||P^\theta_{\bullet|k}) \quad (5)$$

Il serait intéressant de l'étudier conjointement. Nous proposons dans la suite une projection par AFC pour visualiser la conformation de la grille.

L'estimation des paramètres est comparable au TPEM dans le cas gaussien, si ce n'est qu'ici, on doit estimer des paramètres multinomiaux. On prend(Kohonen et al., 2000) un voisinage gaussien (h gaussienne) afin d'obtenir un bon lissage.

8. On rappelle que la divergence de Kullback est toujours positive et s'annule lorsque les deux distributions comparées sont identiques.

9. Si l'on ne considère plus les classes équiprobables, il s'ajoute un terme de régularisation sur la taille des classes qui corrige l'hypothèse implicite de classes équiprobables des moyennes mobiles.

2.3. Méthodes alternatives

Hofmann est à l'origine du PLSA, une méthode probabiliste de réduction de dimension. Il considère les données dyadiques (d_i, m_j) comme multinomiales et suppose d_i et m_j indépendants conditionnellement à une variable x_k inconnue, le facteur latent du PLSA. L'auteur a généralisé dans (Hofmann, 2000) le PLSA en organisant les facteurs en grille topologique. Notre approche suppose des hypothèse multinomiales moins fortes et réduit le nombre des paramètres. Il existe également une méthode de projection à base de multinomiales modélisées par fonction logistique qui correspond à une version pour variables qualitatives du SOM paramétrique de (Bishop et al., 1998).

Les cartes de Sammon(Sammon, 1969) projettent également des données multidimensionnelles sur un plan. Elles ne construisent pas un jeu de vecteurs réduits comme les neurones du SOM et donnent a priori un rendu très différent puisque seules les distances locales y sont conservées. Ici nous tentons d'obtenir une carte conservant au mieux un vision locale mais aussi globale des données grâce à la notion de voisinage. Pour corriger ce défaut, les auteurs de (Demartines and Herault, 1995) ont proposé une méthode ayant des propriétés à la fois des carte de Sammon et carte de Kohonen : elle ne projette plus directement les données mais plutôt une quantification (ex : centres de classe obtenus pas les moyennes mobiles) de celles-ci. L'originalité de cette méthode est de ne plus prendre en compte dans le critère minimisé les distances locales de l'espace de départ mais celui d'arrivé.

3. Méthodes de représentation

Nous décrivons des méthodes de visualisation des données textuelles à partir du modèle de projection par multinomiales. La première est issue de la méthodologie de représentation discrète planaire du SOM alors que la seconde est une projection non discrète sur le plan comparable aux méthodes factorielles.

3.1. Représentations classiques

Une visualisation de la grille de neurones s'effectue à l'aide de la matrice U, le tableau de nombres qui contient les distances entre neurones dans l'espace des données et représenté dans l'espace de sortie ou grille. Il est préconisé d'afficher la matrice U en grisant les cases avec un densité de couleur proportionnelle à la distance inscrite dans le tableau. On peut alors confronter la distance induite artificiellement par la contrainte d'apprentissage en deux dimension avec la distance réelle dans l'espace des données pour juger de la déformation de la carte. Pour des données textuelles, des zones homogènes en thématique apparaissent séparées par des frontières. Une représentation en 3D est aussi envisageable en montrant par exemple l'erreur locale (normalisée). Un document se place de manière naturelle dans la classe où il est le plus probable. La représentation est discrète et ne remplit pas le plan contrairement aux méthodes factorielles.

Il est aussi possible d'effectuer une représentation en calculant une position moyenne sur le plan par $E_{x|m_j,\theta}[x] = \sum_k P(x_k|m_j,\theta)x_k$ et $E_{x|d_i,\theta}[x] = \sum_k P(x_k|d_i,\theta)x_k$, comme dans (Bishop et al., 1998). Néanmoins, cette représentation sensible à la multiplicité des modes peut conduire à une mauvaise interprétation des proximités. Une représentation locale limitée à une seule classe sélectionnée et ses plus proches voisines semble plus pertinent pour une telle projection.

3.2. Représentation factorielle

Nous projetons les centres des multinomiales par analyse factorielle des correspondances (AFC). Il est choisi pour pondération de chacune des classes, la probabilité du facteur multinomial, et pour métrique, l'inverse des probabilités des mots. La projection permet de montrer les dépendances entre facteurs multinomiaux et de juger de la bonne convergence de l'algorithme. Il s'agit d'un complément à la représentation précédente qui montre la divergence plutôt que la distance du χ^2. Il est alors possible de représenter les mots en éléments supplémentaires.

On remarquera que la projection n'est rien d'autre qu'une sorte d'analyse discriminante pour variable qualitative. En effet, en analyse discriminante pour variables continues, la projection maximisant la variance inter projetée est équivalente à l'ACP des centres pondérés par le nombre d'individus présent dans chaque classe. Or, l'estimation du mélange classifiant donne des estimations des poids des classes et valeurs des distributions comparables. Cette représentation enfin est rapide à calculer puisqu'il s'agit de réaliser une AFC sur un nombre réduit d'individus, les facteurs multinomiaux.

Cette représentation est à rapprocher de la méthodologie proposée par Elemento dans (Elemento, 1999) qui préconise une projection factorielle. Il étudie une projection de la carte dont les centres sont vus comme individus supplémentaires, à partir d'une ACP sur l'ensemble des individus. Il propose en outre un initialisation à partir de ces plans pour éviter l'utilisation d'une fonction de voisinage à large variance dès le début de l'algorithme ; en effet, une mauvaise initialisation et un apprentissage mal adapté peuvent conduire à une carte dont la topologie générale respecte peu la distribution des données originales, l'auteur construit une initialisation pré-organisée grâce à l'ACP. Dans notre cas, il est plutôt proposé d'observer après entraînement les dépendances au sens du χ^2 des centres des classes.

4. Indicateurs et expérimentations

D'un point de vue pratique, la carte change d'apparence après chaque nouvel apprentissage si les initialisations sont différentes. Cela provient du fait que la grille de neurones est libre d'effectuer des mouvements (rotations ou symétries) dans l'espace des données, et que des minima locaux peuvent être atteints avant l'optimum. Il faut estimer plusieurs fois les paramètres à partir d'intialisations aléatoires différentes jusqu'à obtenir un résultat satisfaisant. Il faut faire également varier la taille de la carte ainsi que celle du voisinage. C'est un inconvénient de ce genre de méthode à paramètres. Il est éventuellement intéressant d'effectuer ces réglages de manière automatique, mais d'un autre coté, il paraît opportun d'avoir différents points de vue à analyser. Nous définissons des indicateurs qui sont à rapprocher de ceux des méthodes factorielles.

4.1. Indicateurs renseignants les facteurs multinomiaux

La méthode permet de visualiser sur une seule carte une projection globale des données ; or cela induit une déformation inévitable puisque l'on passe d'un espace souvent de l'ordre de plusieurs milliers de dimensions à un espace de dimension deux. Heureusement, l'existence des thématiques en analyse textuelle implique l'existence de dimensions réelles réduites dans les données, et donc une projection porteuse d'information. Il est possible de quantifier les erreurs

globales ou locales comme en analyse factorielle :

- Ici, la contribution relative d'un axe correspond à la probabilité d'un facteur multinomial, c'est-à-dire $P(x_k|\hat{\theta})$;
- Nous proposons de tenir compte de la probabilité $P(m_j|x_k,\hat{\theta})$ d'un mot sur un facteur ainsi que celle d'un document $P(d_i|x_k,\hat{\theta})$ pour renseigner sur l'importance de tel mot ou document sur un facteur multinomial ;
- Par analogie avec Ambroise et Govaert, la conservation topologique du voisinage au niveau des données sur la carte peut s'écrire : $\sum_k \sum_l h(x_k,x_l) KL(P^{\hat{\theta}}_{\bullet|k}||P^{\hat{\theta}}_{\bullet|l})$ où $h(x_k,x_l) = 1$ si $||x_k - x_l||_{\mathcal{R}^2} = 1$ et 0 sinon. On a remplacé la distance euclidienne par la divergence, ce qui est homogène avec le résultat de 2.2 ;
- De même, le critère semblable à la fonction d'erreur des moyennes mobiles obtenu en 2.2 permet de quantifier l'erreur globale de l'approximation ;

4.2. Expériences sur différents corpus

Le vocabulaire $\mathcal{V}$ est construit en gardant environ les 1000 premiers mots les plus fréquents pour chacun des corpus et en supprimant quelques mots outils (une vingtaine) sur-représentés, et considérés comme peu pertinents.

4.2.1. Résumés INRIA

(Morin et al., 2000) ont étudié un corpus d'environ 2000 résumés d'articles de l'INRIA. Les auteurs se servent essentiellement de l'AFC pour cette étude. On projette le même corpus sur une carte comme définie en 2.1. On visualise alors la structure générale du corpus avec les indicateurs proposés. La projection obtenue se trouve en annexe.

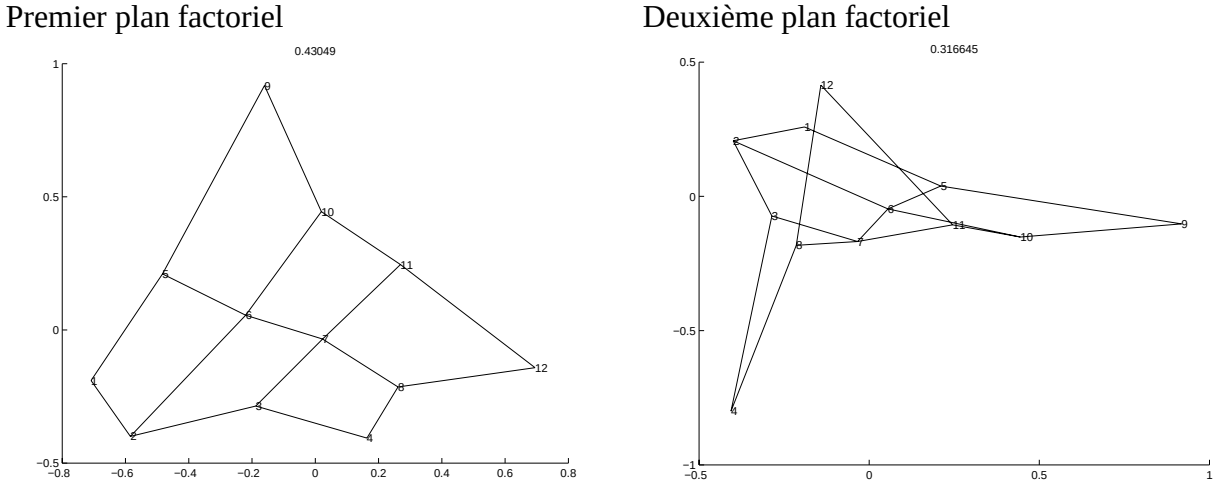

FIG. 2 – *Représentation factorielle (AFC) de la grille de multinomiales (résumés INRIA)*

Nous avons illustré ci-contre un exemple de projection des centres à l'aide de l'AFC. La modèlisation permet de construire des cartes factorielles approchées, ce qui pour un grand nombre de données permet de diminuer fortement les temps de calcul. Ces cartes sont de bons compléments à la matrice U puisqu'elles doivent permettre d'affiner une analyse en segmentant la

carte en thématiques homogènes. On obtient aussi le tableau de coordonnées, contributions et qualités de représentation à consulter en annexe.

La projection sur le le premier plan factoriel n'apporte pas seulement une information visuelle. Si on considère les probabilités de chaque facteur multinomial peu différents, on sait que les contributions de chaque axe sont proportionnelles au carré des coordonnées. Dans ce cas, on peut affirmer, en prenant le centre de gravité (l'origine) centré sur la projection de la carte, que les facteurs de chaque coté de la carte s'opposent d'un point de vue de l'hypothèse d'indépendance étudiée par l'AFC. Ici, par exemple, on peut vérifier que les contributions des facteurs 6-7, proches de l'origine ont une contribution quasiment nulle pour les premiers axes.

En outre, une analyse classique du tableau fourni indique qu'il y a une opposition marquée scindant le tableau en deux partie distinctes. A gauche de la première diagonale se projettent les textes sur les thèmes réseau, système, langage, etc... tandis que sur la droite, on reconnaît un vocabulaire axé davantage sur les thèmes de la modélisation, et des traitements numériques images fixes ou en mouvement, des calcules numériques, etc... La diagonale rassemble les deux grandes thématiques dégagées sur la problématique de l'algorithmique en général. Une étude plus approfondie nécessiterait un retour aux données, d'agrandir la grille de multinomiales, et de commenter un plus grand nombre d'axes factoriels. A ce sujet, on constate sur ces derniers la stabilité du commentaire. Pour un corpus plus complexe, l'étude de ces axes supplémentaires devient inévitable afin de bien segmenter la carte.

4.2.2. Newsgroups WEB

De même, 20000 *newsgroups*(McCallum and Nigam, 1998) servent fréquemment de *benchmark*. L'ensemble de ces *news* a été projeté sur une carte de taille 5*4, comme le nombre de groupes de *news* différents, soit 20. La projection obtenue est en annexe. On ne retrouve pas exactement l'ensemble des *newsgroups*, un par facteur. Certaines thématiques sont en effet peu séparées.

5. Conclusion

Nous avons montré sur un cas particulier qu'une modélisation probabiliste est équivalente à l'usage d'une métrique sous-jacente que nous avons mis en évidence. La divergence est une bonne alternative à la distance euclidienne. Pondérée par la taille des textes, elle permet de prendre en compte des textes de longueur différente, et rappelle le χ^2 pondéré décomposé par l'AFC. La prise en compte du nombre de mots dans les textes est souvent inexistante dans les algorithmes courants ; c'est donc un point important de la projection à base de multinomiales. Les relations entre l'AFC et la méthode que nous proposons sont en cours de développement. on se réferrera à (Priam and Morin, 2001) sur cette question. La segmentation de la matrice U à l'aide de critères statistiques précis est aussi un résultat original.

Il reste à approfondir les expériences empiriques sur un plus grand nombre de corpus afin de mieux évaluer et de perfectionner les méthodes d'apprentissage (difficulté d'un bon entraînement de la grille). Nous envisageons d'améliorer les résultats obtenus, et de les élargir à de plus grandes bases textuelles. Il est par exemple possible d'effectuer un lissage des probabilités d'événement "mot" en ajoutant un modèle a priori sur les paramètres des multinomiales. En outre, la taille de la grille ou la pertinence du vocabulaire employé sont quantifiables par un

test du χ^2. Des expériences avec une distribution binomiale ou Poisson multivariée au lieu de la distribution multinomiale sont aussi envisageables afin de vérifier le bon choix de modèle de texte.

L'usage de la distribution multinomiale permet d'élaborer de nouvelles méthodes de représentation non-linéaire des données qualitatives. Les méthodes probabilistes peu employées jusqu'à présent en visualisation de l'information semblent bien supporter la concurrence des méthodes factorielles grâce à une modélisation pertinente et efficace des données textuelles. la méthode que nous avons obtenue est bien adaptée pour diverses raisons, par exemple (1)-Distribution discrète donnant d'excellents résultats en classification supervisée, (2)-Algorithme performant ne prenant en compte que les valeurs non-nulles de la matrice textuelle, (3)-Complexité algorithmique compétitive puisque si l'on note R, le nombre de composantes non nulles de la matrice de textes, et n le nombre d'itération de l'EM, l'algorithme est seulement de l'ordre de O(nR), (4)-Stockage minimal des données.

Remerciements

Nous remercions Yves LeChevallier (INRIA, Rocquencourt) pour avoir fourni le code C++ d'un logiciel réalisant du SOM. Cela a permis de réaliser le prototype actuel. Grâce à la librairie standard STL, le logiciel a été étendu aux algorithmes présentés et au stockage des grandes matrices creuses textuelles (20000*10000 par exemple), en ne conservant en RAM que les valeurs non nulles. Cela a permis d'accélérer les calculs en prenant en compte uniquement les valeurs non nulles. L'ensemble des graphiques de cet article sont tracés de manière automatique sous MATLAB qui est exécuté par l'implémentation C++.
Nous remercions M. Kerbaol et J.Y. Bansard (INSERM , Rennes) pour leur logiciel BI qui a permis de réaliser des AFCs.

Références

Ambroise C. and Govaert G. (1996). Constrained clustering and Kohonen self-organizing maps. *Journal of Classification*, 13(2):299–313.

Bishop C. M., Svensén M., and Williams C. K. I. (1998). GTM: The Generative Topographic Mapping. *Neural Computation*, 10:215–234.

Demartines P. and Herault J. (1995). CCA: Curvilinear Component Analysis. *15ème Workshop GRETSI, Juan-Les-Pins*.

Droesbeke J.-J., Fichet B., and Tassi P. (1992). *Modèles pour l'analyse des données multidimensionnelles*. Economica.

Elemento O. (1999). Initialisation, convergence, et validation de cartes topologiques de Kohonen, DEA (INRIA, Yves Lechevallier).

Hastie T. and Stuetzle W. (1989). Principal Curves. *Journal of the American Statistical Association*, 84:502–516.

Hofmann T. (2000). Probmap: A Probabilistic Approach for Mapping Large Document Collections. In *IDAJ'2000*.

Kohonen T., Kaski S., Lagus K., Salojrvi J., and et V. Paatero et A. Saarela J. H. (2000). Self-organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11:574–585.

McCallum A. and Nigam K. (1998). A comparison of event models for Naive Bayes text classification. In Press A. editor, *AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48.

Morin A., Kerbaol M., and Bansard J. (2000). Etude des résumés en français des rapports de recherche d'un institut d'informatique publiés de 1989 à 1998. In *JADT'2000*.

Priam R. and Morin A. (2001). Visualisation des données textuelles par analyse factorielle généralisée. *Non publié (voir version courte à EGC'02)*.

Sammon J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18:401–409.

Teuvo Kohonen (1997). *Self-Organizing Maps*. Springer.

ANNEXE - Tableau des indicateurs AFC pour les 6 premiers axes (facteurs multinomiaux estimés à partir des résumés INRIA)

axe	1	2	3	4	5	6
inertie	0,23	0,14	0,13	0,10	0,08	0,05
% inertie	26	17	15	12	9	6

	COORD			CTR (%)			COS2 (%)		
	axe1	axe2	axe3	axe1	axe2	axe3	axe1	axe2	axe3
facteur1	-0,71	-0,19	0,26	31,82	3,55	7,47	60,48	4,34	8,03
facteur2	-0,58	-0,40	0,21	12,21	8,88	2,70	29,01	13,56	3,63
facteur3	-0,19	-0,29	-0,07	0,69	2,47	0,18	3,16	7,32	0,48
facteur4	0,16	-0,41	-0,80	1,39	13,34	58,89	2,92	18,07	70,14
facteur5	-0,48	0,21	0,04	6,33	1,87	0,07	21,45	4,07	0,13
facteur6	-0,22	0,06	-0,05	0,45	0,04	0,04	5,55	0,35	0,25
facteur7	0,02	-0,03	-0,17	0,01	0,02	0,62	0,07	0,14	3,48
facteur8	0,26	-0,21	-0,18	1,19	1,24	1,02	10,29	6,90	5,01
facteur9	-0,16	0,92	-0,10	1,01	51,29	0,73	2,21	72,14	0,90
facteur10	0,02	0,44	-0,15	0,01	9,93	1,32	0,06	30,33	3,54
facteur11	0,27	0,25	-0,11	3,62	4,69	1,00	11,64	9,68	1,81
facteur12	0,69	-0,14	0,41	41,27	2,68	25,95	68,14	2,84	24,23

	COORD			CTR (%)			COS2 (%)		
	Axe4	Axe5	Axe6	Axe4	Axe5	Axe6	Axe4	Axe5	Axe6
facteur1	-0,34	-0,21	0,25	15,35	8,36	16,67	13,52	5,37	7,37
facteur2	0,66	0,30	0,00	33,30	9,80	0,00	36,64	7,87	0,00
facteur3	0,50	0,31	-0,22	10,59	5,71	4,00	22,58	8,88	4,29
facteur4	-0,22	0,11	0,07	5,49	1,81	1,03	5,36	1,29	0,50
facteur5	-0,44	-0,06	-0,76	11,56	0,27	67,86	18,14	0,31	53,46
facteur6	0,15	0,01	-0,26	0,46	0,00	2,73	2,61	0,01	7,85
facteur7	0,25	-0,16	-0,09	1,65	0,91	0,45	7,56	3,05	1,05
facteur8	-0,15	-0,02	0,01	0,85	0,01	0,01	3,39	0,04	0,02
facteur9	-0,09	0,47	0,20	0,74	25,09	7,03	0,75	18,55	3,58
facteur10	0,19	-0,20	-0,01	2,56	3,76	0,01	5,64	6,03	0,01
facteur11	0,36	-0,53	0,03	14,05	41,13	0,20	20,91	44,65	0,15
facteur12	-0,14	0,11	0,00	3,41	3,15	0,00	2,61	1,76	0,00

ANNEXE - Exemple de grille des 2000 résumés INRIA

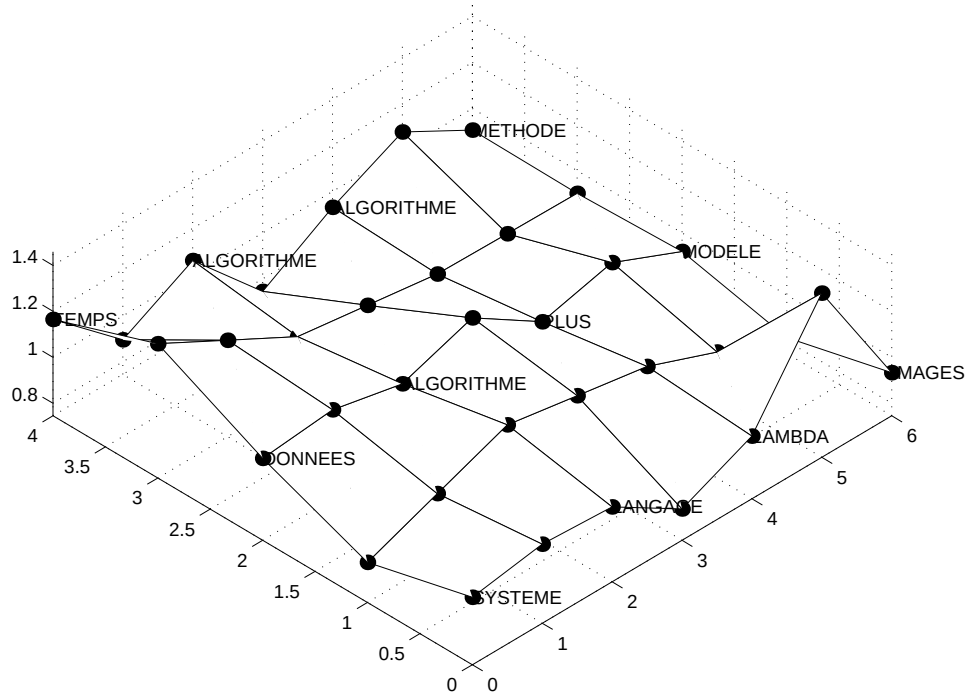

FIG. 3 – *matrice U des résumés pour la grille ci-dessous*

SYSTEME	LANGAGE	LAMBDA	IMAGES
DONNEES	SYSTEME	TYPES	METHODE
LANGAGE	SEMANTIQUE	SYSTEME	DEUX
APPLICATIONS	TYPES	ORDRE	POINTS
RAPPORT	LOGIQUE	PLUS	PLUS
OBJETS	PREUVE	CALCUL	IMAGE
PROGRAMMATION	PLUS	TERMES	MOUVEMENT
PLUS	CALCUL	TYPE	PROBLEME
MODELE	LANGAGES	METHODE	PARTIR
SYSTEMES	PROGRAMMATION	ES	MODELE
DONNEES	ALGORITHME	PLUS	MODELE
MEMOIRE	ETRE	CLASSIFICATION	METHODE
ETRE	SYSTEMES	ALGORITHMES	ALGORITHME
EXECUTION	PEUVENT	METHODE	DONNEES
ALGORITHME	PLUS	CRITERES	RAPPORT
PEUT	ENSEMBLE	SYSTEMES	PROBLEME
ARTICLE	ENTRE	THEORIE	MODELES
ALGORITHMES	ALGORITHMES	MODELES	IMAGES
COHERENCE	REGLES	ORDRE	DEUX
SYSTEMES	ORDRE	COMME	RESULTATS
TEMPS	ALGORITHME	ALGORITHME	METHODE
RESEAUX	RESEAUX	ANALYSE	PROBLEME
PLUS	TEMPS	ALGORITHMES	EQUATIONS
ORDONNANCEMENT	PLUS	PROBLEME	ORDRE
SYSTEME	NOMBRE	METHODE	PROBLEMES
RESEAU	PROCESSUS	PLUS	RAPPORT
PROBLEME	DEUX	FONCTIONS	METHODES
ATTENTE	ETRE	PROBLEMES	CONDITIONS
NOMBRE	PROBLEME	CALCUL	EQUATION
DEUX	ALGORITHMES	RESULTATS	FINIS

ANNEXE - Exemple de grille des 20000 newsgroups WEB

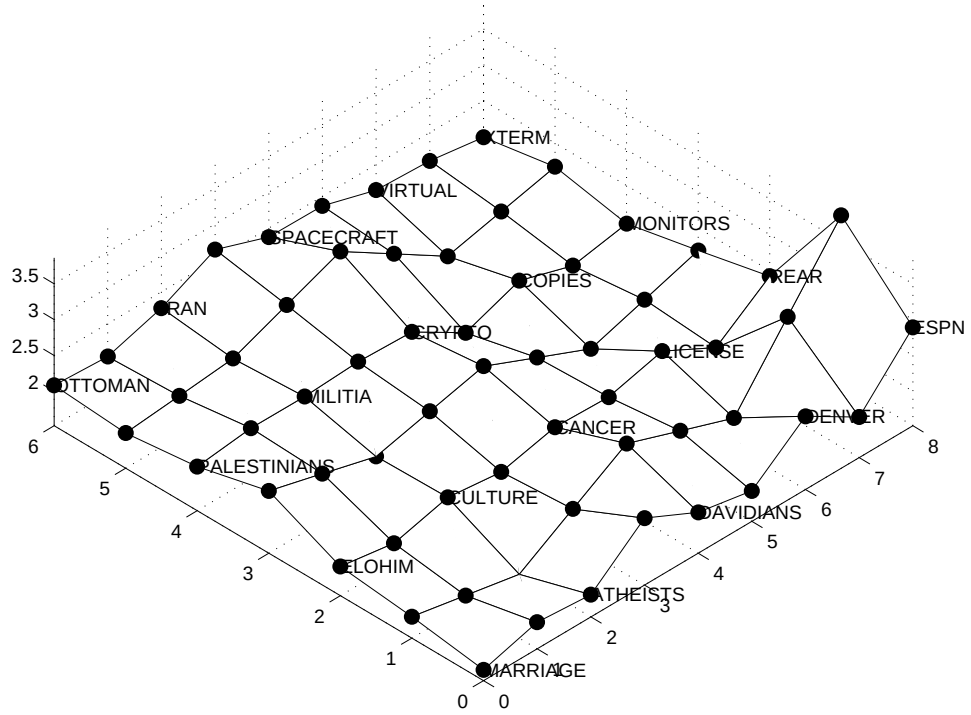

FIG. 4 – *matrice U des news pour la grille ci-dessous*

MARRIAGE	ATHEISTS	DAVIDIANS	DENVER	ESPN
APARTMENT	SUBJECTIVE	CULT	SEMI	STATS
GRACE	GODS	TEAR	JOKE	PLAYOFFS
MATTHEW	ASSUMPTION	TEXAS	BUSH	BRAVES
SUMGAIT	JUDGEMENT	ASSAULT	HOSPITAL	PITCHER
LUKE	LOGICAL	SUICIDE	ROCK	CHICAGO
VERSE	PERSPECTIVE	RAID	TRAINING	DETROIT
ETERNAL	ASSERTION	TRIAL	BRIAN	PITCHING
TESTAMENT	BELIEVED	FEDS	HITTING	PITTSBURGH
REVELATION	TOPIC	KNOCK	SHOTS	MORRIS
ELOHIM	CULTURE	CANCER	LICENSE	REAR
MORMON	MALE	MEDICINE	OWNERS	RIDING
UNTO	EXAMPLES	DIET	SALES	FORD
GODS	MINORITY	PATIENT	DETECTOR	TIRES
DOCTRINE	DECISIONS	BRAIN	PASSED	WIRING
MOVEMENT	HOMOSEXUALS	CANDIDA	TAXES	BIKES
BIBLICAL	IGNORANT	SYMPTOMS	OWNER	BRAKE
LEADERS	SCHOOLS	CHRONIC	WAITING	HELMET
JOSEPH	THREAT	DOCTORS	ACCIDENT	NEUTRAL
VERSES	RATIONAL	LEVELS	RIDING	HONDA
PALESTINIANS	MILITIA	CRYPTO	COPIES	MONITORS
PALESTINIAN	BEAR	ENCRYPTED	ENGINES	CHANNEL
ISLAMIC	CONSTITUTIONAL	PHONES	LETTERS	ELECTRONICS
LEBANON	SUPREME	COMMUNICATIONS	STUDENT	VOLTAGE
SERBS	WEAPON	PROPOSAL	RECORDS	SONY
CIVILIANS	NUCLEAR	SCHEME	SALES	LASER
BOSNIA	LIBERTARIAN	AGENCIES	COMPONENTS	STEREO
JERUSALEM	ORGANIZED	AGENCY	STORE	PORTABLE
NAZIS	FIGHTING	CRYPTOGRAPHY	ANALYSIS	ITEMS
NATIONS	PAPERS	EXPORT	CORP	WARRANTY
OTTOMAN	IRAN	SPACECRAFT	VIRTUAL	XTERM
GREECE	JOIN	LUNAR	PLANE	BIOS
AZERI	JUNE	MARS	OBJECTS	POSTSCRIPT
EMPIRE	REGION	VENUS	CRYPTOGRAPHY	SIMMS
PARAGRAPH	RUSSIA	LARSON	FUNCTIONS	OPENWINDOWS
REPUBLIC	TROOPS	SATELLITE	DISTANCE	CACHE
KARABAKH	GREECE	PROBE	TELEPHONE	BYTE
MASSACRE	NUCLEAR	ASTRONOMY	PROCESSING	SHELL
KURDS	TRAINING	TEMPERATURE	SESSION	PROCESSOR
VILLAGE	POTENTIAL	PROPULSION	FREQUENCY	TRANSFER