ANR CLASEL

Mohamed Nadif[1], **Gérard Govaert**[2]

[1] LIPADE, Université Paris Descartes, France
[2] Heudiasyc, Université de Technologie de Compiègne, France

Introduction

- Types d'algorithmes de classification croisée
- Intérêts
- Cas des données continues
- Approche modèle de mélange

Livrables

- Modèles pour la classification croisée
  - Livrable 1.1 : Etat de l'art
  - Livrable 1.2 : Modèles pour les données continues
- Sélection de modèle
  - Livrable 2.3 : Aspects asymptotiques

Travaux en cours et programme

Données

- matrix $\mathbf{x} = (x_{ij})$
- $i \in I$ ensemble de n lignes
- $j \in J$ ensemble de d colonnes

Partition z de I en g classes

- $\mathbf{z} = (z_{ik})$: une matrice de classification de taille $(n \times g)$
- $z_{ik} = 1$ si $i \in k$ème classe and $z_{ik} = 0$ sinon

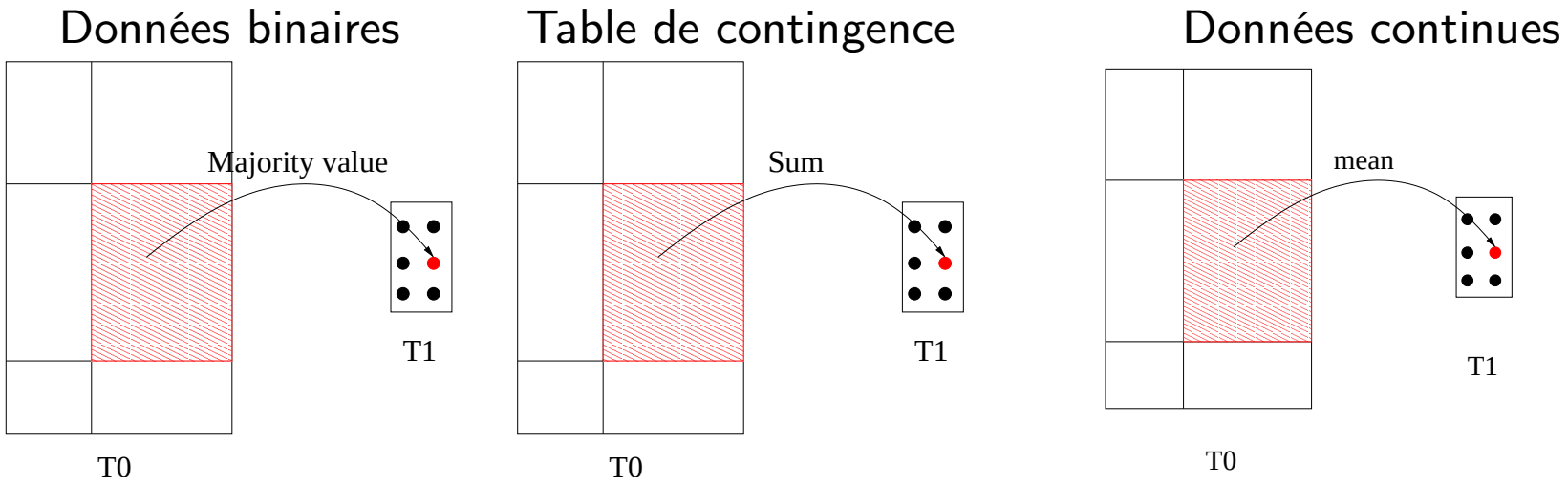
3	0	0	1
2	0	1	0
3	0	0	1
2	0	1	0
1	1	0	0

Partition w de J en m classes

- $\mathbf{w} = (w_{j\ell})$: une matrice de classification de taille $(d \times m)$
- $w_{j\ell} = 1$ si $j \in \ell$ème classe et $w_{j\ell} = 0$ sinon

Bloc $k\ell$ défini par i, j tel que $z_{ik} w_{j\ell} = 1$

Principe général



Critères

Données	$\mu_{k\ell}$	Critère $W(\mathbf{z}, \mathbf{w}, \mu)$
Binaires	Mode	$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} \lvert x_{ij} - \mu_{k\ell} \rvert$
Table de contingence	Somme	$\chi^2(\mathbf{z}, \mathbf{w}) = K \sum_{k,\ell} \frac{(f_{k\ell} - f_{k.} f_{.\ell})^2}{f_{k.} f_{.\ell}}$
Continues	Moyenne	$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - \boldsymbol{\mu}_{k\ell})^2 = \lvert\lvert \mathbf{x} - \mathbf{z}\boldsymbol{\mu}\mathbf{w}^T \rvert\rvert^2$

Complémentaires aux méthodes factorielles

- ACP, AC, etc.

Réduction de la taille des données

- Offre un résumé (de même nature) de la table initiale

Méthodes capables de gérer des données de grande taille

- Moins de calcul par l'utilisation de matrices de taille réduite
- Remédie à la grande dimension et la *sparsité*

Applications

- *Classification de documents* : classification des documents et des mots simultanément
- *Données des biopuces* : classification des gènes et des tissus simultanément
- *Systèmes de recommandation*

Minimisation du critère $W(\mathbf{z}, \mathbf{w}, \boldsymbol{\mu}) = ||\mathbf{x} - \mathbf{z}\boldsymbol{\mu}\mathbf{w}^T||^2$

Double k-means

(a) Recherche de $\boldsymbol{\mu} = (\mu_{k\ell})$ et $\mathbf{z}$ en minimisant $W(\mathbf{z}, \boldsymbol{\mu}|\mathbf{w}) = \sum_k \sum_{j,\ell} w_{j\ell}(x_{ij} - \mu_{k\ell})^2$

(b) Recherche de $\boldsymbol{\mu} = (\mu_{k\ell})$ et $\mathbf{w}$ en minimisant $W(\mathbf{w}, \boldsymbol{\mu}|\mathbf{z}) = \sum_\ell \sum_{i,k} z_{ik}(x_{ij} - \mu_{k\ell})^2$

Algorithme *Croeuc* : Govaert (1983)

Working on intermediate $(n \times m)$ matrix $\mathbf{u} = (u_{i\ell})$ et $(g \times d)$ matrix $\mathbf{v} = (v_{kj})$

(a) minimisation de $W(\mathbf{z}, \boldsymbol{\mu}|\mathbf{w}, \mathbf{u})$ où $u_{i\ell} = \sum_j w_{j\ell}x_{ij}/w_\ell$ $(w_\ell = \sum_j w_{j\ell})$

(a.1) *k-means* sur $\mathbf{u}$ et on obtient $\mathbf{z}$ et $\boldsymbol{\mu}$

(b) minimisation de $W(\mathbf{w}, \boldsymbol{\mu}|\mathbf{z}, \mathbf{v})$ où $v_{kj} = \sum_i z_{ik}x_{ij}/z_k$ $(z_k = \sum_i z_{ik})$

(b.1) *k-means* sur $\mathbf{v}$ et on obtient $\mathbf{w}$ et $\boldsymbol{\mu}$

Croeuc vs Double k-means et autres variantes

- Résultats équivalents
- *Croeuc* est plus rapide

Plusieurs problèmes avec ce critère (prop. varia.) : approche *Latent Block Model*

Latent block model, Govaert and Nadif (2003)

Définition du modèle de mélange sur $I \times J$

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{\mathbf{u} \in U} P(\mathbf{u}) f(\mathbf{x}|\mathbf{u}; \boldsymbol{\alpha})$$

où U est l'ensemble de toutes les partitions de $I \times J$

Hypothèses supplémentaires pour considérer la classification croisée

- $\mathbf{u} = \mathbf{z} \times \mathbf{w}$
- $f(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\alpha}) = \prod_{i,j} \varphi(x_{ij}; \boldsymbol{\alpha}_{z_i, w_j})$ où $\varphi(., \boldsymbol{\alpha})$ est pdf sur $\mathbb{R}$

Latent block model

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} \underbrace{\prod_{i,k} \pi_k^{z_{ik}}}_{prop.} \underbrace{\prod_{j,\ell} \rho_\ell^{w_{j\ell}}}_{prop.} \underbrace{\prod_{i,j,k,\ell} \varphi(x_{ij}; \boldsymbol{\alpha}_{k\ell})^{z_{ik} w_{j\ell}}}_{dens.}$$

où $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \rho_1, \ldots, \rho_m, \boldsymbol{\alpha}_{11}, \ldots, \boldsymbol{\alpha}_{gm})$

Types de données : différentes approches, ML, CML et floue

Bernoulli latent block model, **Govaert and Nadif (2003, 2005a, 2008)**

- Données binaires
- φ distribution de Bernoulli $\mathcal{B}(\alpha_{k\ell})$

Poisson latent block model, **Govaert and Nadif (2005b)**

- Table de contingence
- φ distribution de Poisson $\mathcal{P}(\mu_i \nu_j \alpha_{k\ell})$
  - μ_i et ν_j effets de i et de j
  - $\alpha_{k\ell}$ effet du bloc $k\ell$.

Avantages

- Modèles plus parcimonieux qu'un modèle de mélange classique sur I et sur J
- Critères plus riches et algorithmes de type *Block EM* plus efficaces pour la classification croisée

A faire

- Extension au cas continu possible
- Sélection de modèle

Historique

- Travaux depuis 1970

Difficultés

- Approches et objectifs différents

Domaines

- Statistique
- Analyse de données textuelles
- Filtrage collaborative
- Factorisation de matrice non-négative
- Bioinformatique
- Approche modèle de mélange

Résultat

- Un rapport avec 124 références
- Une soumission dans une revue prévue avant fin décembre

Densité du mélange

- $f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_k \pi_k \varphi_k(\mathbf{x}_i;\mathbf{w},\boldsymbol{\alpha})$ où
  - $\varphi_k(\mathbf{x}_i;\mathbf{w},\boldsymbol{\alpha}) = \prod_{j,\ell} \left(\frac{1}{\sqrt{2\pi\sigma^2_{k\ell}}} e^{-\frac{1}{2\sigma^2_{k\ell}}(x_{ij}-\mu_{k\ell})^2} \right)^{w_{j\ell}}$
  - $\boldsymbol{\theta} = (\pi_1,\ldots,\pi_g,\mathbf{w},\boldsymbol{\alpha})$

$$\boldsymbol{\alpha} = \begin{pmatrix} (\mu_{k1},\sigma^2_{k1}) & \ldots & (\mu_{km},\sigma^2_{km}) \\ \ldots & \ldots & \ldots \\ (\mu_{g1},\sigma^2_{g1}) & \ldots & (\mu_{gm},\sigma^2_{gm}) \end{pmatrix}$$

Lien entre ce modèle et le modèle gaussien sphérique

$$\mu_k = (\mu_{k1},\ldots,\mu_{k1},\mu_{k2},\ldots,\mu_{k2},\ldots,\mu_{km},\ldots,\mu_{km})^T$$

$$\Sigma_k = \begin{pmatrix} \sigma^2_{k1} & & & & & & \\ & \ldots & & & & & \\ & & \sigma^2_{k1} & & & & \\ & & & \ldots & & & \\ & & & & \sigma^2_{km} & & \\ & & & & & \ldots & \\ & & & & & & \sigma^2_{km} \end{pmatrix}$$

Caractéristiques du modèle

- Nombre de paramètres égal $g + (g \times m) + (g \times m)$ au lieu de $g + (g \times d) + (g \times d)$.
- Parcimonie très utile lorsque n<d

Algorithme GEM

- E-step : Calcul de $s_{ik}^{(c)} = P(z_{ik} = 1|\mathbf{x}, \boldsymbol{\theta}^{(c)}) \propto \pi_k^{(c)} \varphi_k(\mathbf{x}_i; \mathbf{w}^{(c)}, \boldsymbol{\alpha}^{(c)})$
- M-step : Maximisation de $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)}) = \sum_{i,k} s_{ik}^{(c)} \log(\pi_k \varphi_k(\mathbf{x}_i; \mathbf{w}, \alpha))$

Algorithme CGEM

- Lorsque toutes les variances sont supposées égales 1) la maximisation de la vraisemblance classifiante est équivalente à la minimisation de $||\mathbf{x} - \mathbf{z}\boldsymbol{\mu}\mathbf{w}^T||^2$ 2) La version CGEM est équivalente à *Croeuc*

Résultats

- Une publication dans CAP'2009
- Une publication dans ICMLA'2010
- Préparation d'une version étendue

Sélection de modèle

- Extension du critère BIC : $BIC(g, m) = L - d\log(m) - \frac{\nu}{2}\log(nd)$
- CAP'2009

Thèse, Heudiasyc

- Expression du critère de BIC au modèle *Latent block model* par une approximation variationnelle étendue
- Sujet de thèse de Aurore Lomet (financée par le ministère)

Post-doc, Heudiasyc

- Principe du bootstrap pour corriger l'erreur de l'estimateur de la vraisemblance ou la vraisemblance classifiante
- Modèles de mélange Gaussiens simples
- Plusieurs critères bootstrap expérimentés

Travaux dans les 6 mois à venir

Heudiasyc & LIPADE

- Plusieurs travaux en cours de finalisation
- Etude des liens entre le modèle proposé et la tri-factorisation de matrice non négative (post-doc, LIPADE, 01/09/10)
- Visualisation des données par GTM (en cours de soumission)

Heudiasyc

- Version Stochastique de *Block EM*
- Sélection de modèle

LIPADE

- Données manquantes : Algorithmes qui fonctionnent
- Bagging dans la classification croisée (ECML/PKDD 2010)