Classification partiellement supervisée

Gérard Govaert

HEUDIASYC, UMR CNRS 6599

Université de Technologie de Compiègne

gerard.govaert@utc.fr

21 Mars 2003

Introduction

- Point de départ : problème du Cetim, puis problème d'indexation d'images (études en cours)
  - Recherche de défauts sur des cuves
  - Émissions acoustiques classées en 3 classes : pas de défaut, défaut bénin, défaut grave
  - Quelque fois hésitation entre les 2 types de défaut
  - Comment prendre en compte cette information ?
  - Extension du cadre habituel des données d'apprentissage

- Problème d'apprentissage :
  - Individus caractérisé par p mesures
  - Classification : « apprendre »une fonction : $\mathbb{R}^p \to \{1, \ldots, g\}$
  - Régression : « apprendre »une fonction : $\mathbb{R}^p \to \mathbb{R}$
- Données disponibles :
  - Individus caractérisés par p mesures + étiquette : Apprentissage Supervisée ou discrimination
  - Individus caractérisés par p mesures : Apprentissage non supervisée ou clustering

- Existence de situations mixtes :
  - ➤ Grosses masses de données non étiquetées peu coûteuses
  - ➤ Nombre réduit de données étiquetées coûteuses (intervention humaine)
- Quelques exemples
  - Données satellitaires
  - Données textuelles (Web)
  - Indexation d'images
- Problème :
  - ➤ Discrimination : on perd les individus non étiquetés
  - ➤ Clustering : on perd les labels des individus étiquetés
- Idée : utiliser toute l'information → Apprentissage semi supervisé

- Deux points de vue :
  - Clustering + ajout d'information supplémentaire provenant des individus étiquetés
  - Discrimination + ajout d'information supplémentaire provenant des individus non étiquetés
- Problème ancien
  - Titterington (1976), O'Neil (1978), Ganesalingam et Mclachlan (1978)

  mais qui fait actuellement l'objet de nombreux développements dans le monde de l'apprentissage
  - Classification de textes (Web, email)(Nigam et al., 2000)

- Application à des méthodes variées
  - Modèle de mélange et analyse discriminante linéaire (O'Neil, 1978, Ganesalingam et Mclachlan, 1978)
  - Estimation de densité par la méthode des noyaux (Murray et Titterington, 1978)
  - Approche bayésienne (Titterington, 1976)
  - Discrimination logistique (Anderson, 1979)
  - SVM
  - Co-training (Blum et Mitchell, 1998)
  - Clustering avec algorithme génétique (Demeriz et al., 2002)
  - Classification hiérarchique (Larsen et al., 2001)
  - Approches fonction de croyance

- Extension : étiquetage partiel
  - L'individu n'appartient pas à la classe C mais il appartient à la classe A ou à la classe B
  - Exemple dans le domaine médical : un médecin est certain que son patient n'est pas atteint de la maladie C mais hésite entre la maladie A ou la maladie B
  - Exemple du Cetim
  - Modélisation hiérarchique par modèle de mélange
  - Apprentissage partiellement supervisé

- Suivant les approches retenus, les données non étiquetées peuvant apporter plus ou moins d'information supplémentaire (aucune dans les modèles logistiques pures)
- Cadre retenu ici
  - Lié au travail avec le Cetim
  - Classification s'appuyant sur le modèle de mélange (approche générative)
  - Indépendance entre le fait de connaître ou non la classe et la valeur $\mathbf{x}$

Plan

- Cadre général de l'apprentissage
- Mélange fini de lois de probabilité
- Apprentissage semi supervisé avec le modèle de mélange
- Apprentissage partiellement supervisé avec le modèle de mélange
- Application à la fouille d'images
- Conclusion

Cadre général de l'apprentissage

▶ Données

- Observations $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ n vecteurs de $\mathbb{R}^d$ (échantillon iid)
- Existence de g classes : matrice de classification $\mathcal{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ n vecteurs de $\{0,1\}^g$ vérifiant
  $z_{ik} \in \{0,1\}$ et $\forall i, \sum_{k=1}^{g} z_{ik} = 1$
- Seulement une partie de $\mathcal{Z}$ est connue
- Données complètes : $(\mathcal{X}, \mathcal{Z})$

$\mathcal{X}$				$\mathcal{Z}$		
3.5	2.3	0.3	4.2	0	0	1
2.2	1.4	2.9	1.3	0	1	0
4.2	1.7	2.2	1.1	0	0	1
2.5	2.3	0.3	4.2	0	0	1
9.2	2.4	2.9	1.3	1	0	0
6.2	1.2	2.2	1.1	1	0	0

- **Des problèmes distincts**
  - Apprentissage supervisé
  - Apprentissage non supervisé
  - Apprentissage semi-supervisé
  - Apprentissage partiellement supervisé

▶ Apprentissage supervisé

- Données :
  - ➤ Un ensemble d'apprentissage $(\mathcal{X}, \mathcal{Z})$
  - ➤ Une observation $\mathbf{x}$

3.5	2.3	0.3	4.2	0	0	1
2.2	1.4	2.9	1.3	0	1	0
4.2	1.7	2.2	1.1	0	0	1
2.5	2.3	0.3	4.2	0	0	1
9.2	2.4	2.9	1.3	1	0	0
6.2	1.2	2.2	1.1	1	0	0
4.4	2.4	2.1	1.0	?	?	?

- Problème : estimer $\mathbf{z}$ la classe de $\mathbf{x}$ à partir de $(\mathcal{X}, \mathcal{Z})$

▶ Apprentissage non-supervisé (Clustering)

- Données : $\mathcal{X}$

$\mathcal{X}$				$\mathcal{Z}$		
3.5	2.3	0.3	4.2	?	?	?
2.2	1.4	2.9	1.3	?	?	?
4.2	1.7	2.2	1.1	?	?	?
2.5	2.3	0.3	4.2	?	?	?
9.2	2.4	2.9	1.3	?	?	?
6.2	1.2	2.2	1.1	?	?	?

- Problème : estimer la matrice de classification $\mathcal{Z}$ à partir de $\mathcal{X}$

▶ Apprentissage semi-supervisé

- Données : $(\mathcal{X}, \mathcal{Z}_1) = ((\mathbf{x}_1, \mathbf{z}_1), (\mathbf{x}_2, \mathbf{z}_2), \ldots (\mathbf{x}_m, \mathbf{z}_m), \mathbf{x}_{m+1}, \ldots, \mathbf{x}_n)$

$\mathcal{X}$				$\mathcal{Z}$		
3.5	2.3	0.3	4.2	0	0	1
4.2	1.7	2.2	1.1	1	0	0
6.2	1.2	2.2	1.1	0	1	0
2.5	2.3	0.3	4.2	?	?	?
9.2	2.4	2.9	1.3	?	?	?
2.2	1.4	2.9	1.3	?	?	?
7.2	2.9	7.9	1.6	?	?	?

- $\mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2)$
- Problème : estimer $\mathcal{Z}_2 = (\mathbf{z}_{m+1}, \ldots, \mathbf{z}_n)$ à partir $(\mathcal{X}, \mathcal{Z}_1)$

▶ Apprentissage partiellement supervisé

- Données : $(\mathcal{X}, \mathcal{Z}_1)$

3.5	2.3	0.3	4.2	0	0	1
2.2	1.4	2.9	1.3	0	?	?
4.2	1.7	2.2	1.1	0	0	1
2.5	2.3	0.3	4.2	?	?	?
4.3	2.9	6.2	4.9	?	0	?
9.2	2.4	2.9	1.3	?	?	?
6.2	1.2	2.2	1.1	1	0	0

- $\mathcal{Z} = (\mathcal{Z}_1, \mathcal{Z}_2)$
- Problème : estimer $\mathcal{Z}_2$ à partir $(\mathcal{X}, \mathcal{Z}_1)$

Mélange fini de lois de probabilité

Modèle de référence dans cette présentation

- **Principe :**

Connaissant

- les proportions $\pi_1, \ldots, \pi_g$
- les distributions de chaque classe

les données sont générées suivant le mécanisme suivant :

- $\mathbf{z}$: chaque individu est rangé dans une classe suivant la loi multinomiale de paramètres $\pi_1, \ldots, \pi_g$
- $\mathbf{x}$: chaque $\mathbf{x}_i$ suit la loi de probabilité associée à la classe à laquelle il appartient

▶ Loi de probabilité

Plus formellement, cela revient à supposer que les $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sont issus d'un vecteur aléatoire de densité

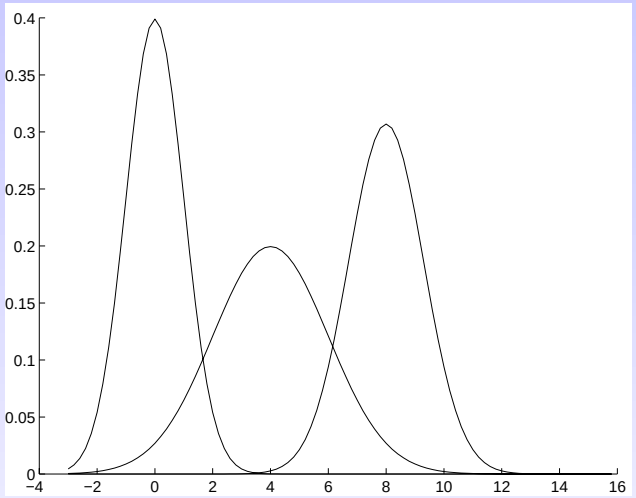
$$p(\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \varphi_k(\mathbf{x}; \boldsymbol{\alpha}_k)$$

où

- $\varphi_k \in$ famille de distributions de $\mathbb{R}^d$
- $\boldsymbol{\theta} = (\Pi, \boldsymbol{\alpha})$ paramètre du modèle :
  - $\Pi = (\pi_1, \ldots, \pi_g)$ proportions du mélange
  - $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ paramètres des densités de chaque classe

▶ Exemples

(a) Les composantes



(b) Le mélange

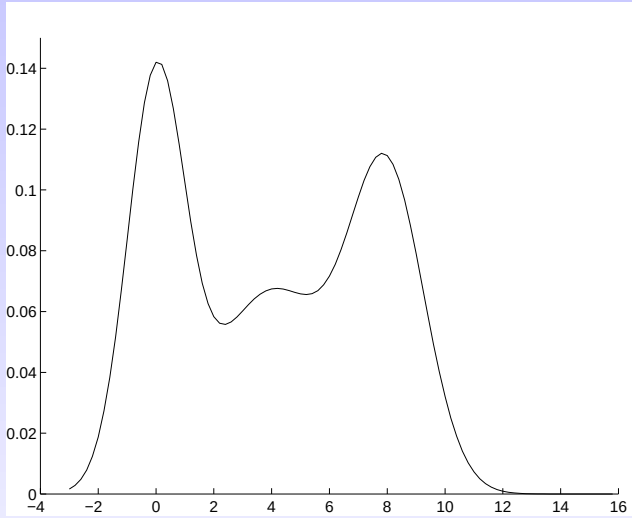


FIG. 1 – Mélange gaussien dans $\mathbb{R}$

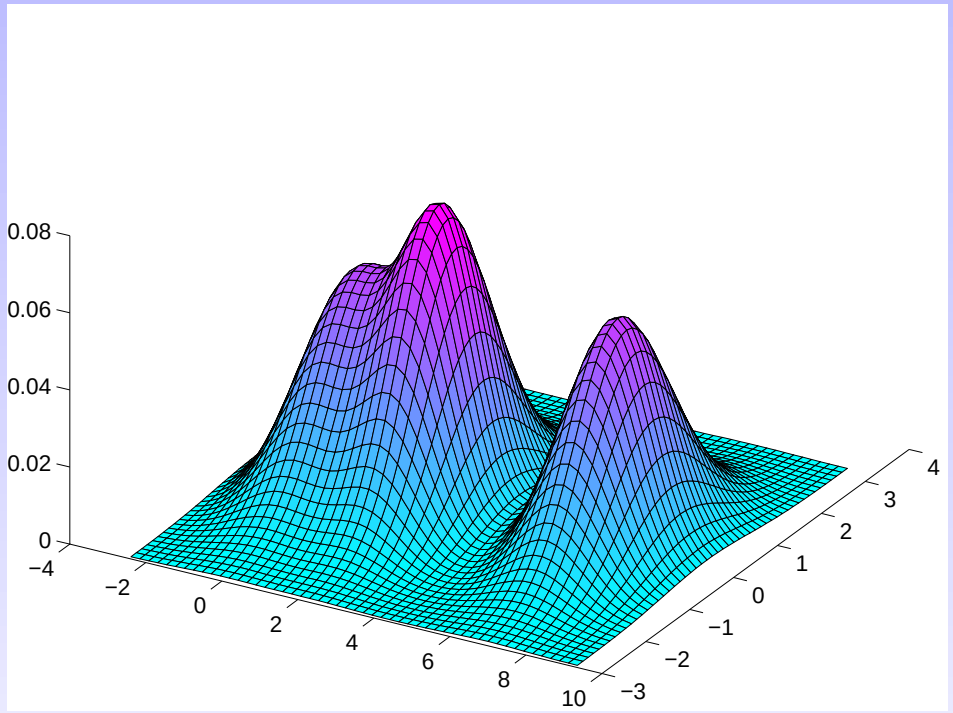


FIG. 2 – Mélange gaussien dans $\mathbb{R}^2$

▶ Estimation des paramètres

- Approches variées (i.e. Karl Pearson 1894, méthode des moments)
- Une méthode largement utilisée : maximiser la log-vraisemblance

$$L(\boldsymbol{\theta}; \mathcal{X}) = \ln \left(\prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta}) \right) = \sum_{i=1}^{n} \ln \left(\sum_{k} \pi_k \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k) \right)$$

- Pas de solution analytique
- Solution classique : algorithme EM (Dempster, Laird et Rubin, 1977)

▶ Principe de l'algorithme EM

- Données complétées $\mathcal{Y}$
- Maximisation de la vraisemblance complétée $L(\boldsymbol{\theta};\mathcal{Y})$ simple
- $f(\mathcal{X};\boldsymbol{\theta}) = f(\mathcal{Y};\boldsymbol{\theta})/p(\mathcal{Y}/\mathcal{X},\boldsymbol{\theta}) \Rightarrow L(\boldsymbol{\theta};\mathcal{X}) = L(\boldsymbol{\theta};\mathcal{Y}) - \log p(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta})$
- Espérance cond. à $\mathcal{X}$ et à une estimation courante $\boldsymbol{\theta}^c$:

$$L(\boldsymbol{\theta};\mathcal{X}) = \underbrace{E(L(\boldsymbol{\theta};\mathcal{Y})|\mathbf{x},\boldsymbol{\theta}^c)}_{Q(\boldsymbol{\theta},\boldsymbol{\theta}^c)} - \underbrace{E(p(\mathcal{Y}|\mathcal{X},\boldsymbol{\theta}))}_{H(\boldsymbol{\theta},\boldsymbol{\theta}^c)}$$

- En prenant $\boldsymbol{\theta}^{c+1} = \arg\max Q(.,\boldsymbol{\theta}^c)$, on obtient
  - $Q(\boldsymbol{\theta}^{c+1},\boldsymbol{\theta}^c) \geq Q(\boldsymbol{\theta}^c,\boldsymbol{\theta}^c)$
  - Mais $H(\boldsymbol{\theta}^{c+1},\boldsymbol{\theta}^c) \leq H(\boldsymbol{\theta}^c,\boldsymbol{\theta}^c)$ (Inégalité de Jensen) et donc

$$L(\boldsymbol{\theta}^{c+1};\mathbf{x}) \geq L(\boldsymbol{\theta}^c;\mathbf{x})$$

- Mélange :
  - Données complétées $\mathcal{Y} = (\mathcal{X}, \mathcal{Z})$

$$L(\boldsymbol{\theta}; \mathcal{X}, \mathcal{Z}) = \sum_i \sum_k z_{ik} \log\{\pi_k \varphi_k(\mathbf{x}_i, \boldsymbol{\alpha}_k)\}$$

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^c) = \sum_i \sum_k P(z_{ik} = 1 | \mathbf{x}_i, \boldsymbol{\theta}^c) \log\{\pi_k \varphi_k(\mathbf{x}_i, \boldsymbol{\alpha}_k)\}$$

▶ L'algorithme EM

- étape 0 : solution initiale $(\pi^0, \boldsymbol{\alpha}^0)$
- étape E : calcul de $t_{ik}^c = P(\mathbf{x}_i \in P_k | \mathbf{x}_i, \pi^c, \boldsymbol{\alpha}^c)$:

$$t_{ik}^c = \frac{\pi_k^m \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k^c)}{\sum_{\ell=1}^g \pi_\ell^c \varphi_\ell(\mathbf{x}_i; \boldsymbol{\alpha}_\ell^c)}$$

- étape M : maximisation de la vraisemblance cond. aux t_{ik}^c :
  - $\pi_k^{c+1} = \frac{1}{n} \sum_{i=1}^n t_{ik}^c$
  - $\boldsymbol{\alpha}_k^{c+1}$: résolution des équations de vraisemblance

▶ Exemple du mélange gaussien

- $\boldsymbol{\alpha}_k = (\mu_k, \Sigma_k)$
- $\mu_k^{c+1} = \frac{1}{\sum_{i=1}^n t_{ik}^c} \sum_{i=1}^n t_{ik}^c \mathbf{x}_i$
- $\Sigma_k^{c+1} = \frac{1}{\sum_{i=1}^n t_{ik}^c} \sum_{i=1}^n t_{ik}^c (\mathbf{x}_i - \mu_k^{c+1})(\mathbf{x}_i - \mu_k^{c+1})'$

▶ Propriétés de l'algorithme EM

- En général très simple à mettre en place
- Optimisation locale de la vraisemblance
- Bon comportement pratique
- Peut être lent dans certains cas : classes très mélangées par exemple

▶ Utilisation du mélange en classification

- Approche probabiliste qui paraît la plus naturelle :
  chaque classe est modélisée par une distribution de probabilité

$$i \in k \Longrightarrow \mathbf{x}_i \sim \mathcal{L}_k$$

- Références : Scott and Symons (1971), Marriott (1974), Symons (1981), McLachlan and Basford (1988)
- Pb posé : Retrouver le composant dont est issu chaque $\mathbf{x}_i$: trouver $\mathcal{Z}$ à partir de $\mathcal{X}$
- Le nombre de classes k sera généralement supposé connu

- Deux approches possibles :
  - **Approche vraisemblance**
    - Estimation de $\boldsymbol{\theta}$ avec EM : $\widehat{\boldsymbol{\theta}}$
    - Estimation de $\mathcal{Z}$: MAP$(t_{ik}(\widehat{\boldsymbol{\theta}}))$
  - **Approche vraisemblance classifiante**
    - Estimation simultanée de $\boldsymbol{\theta}$ et de $\mathcal{Z}$ par maximisation de la vraisemblance complétée (souvent appelée ici vraisemblance classifiante)

$$L(\boldsymbol{\theta}; \mathcal{X}, \mathcal{Z}) = \sum_i \sum_k z_{ik} \log\{\pi_k \varphi_k(\mathbf{x}_i, \boldsymbol{\alpha}_k)\}$$

à l'aide de l'algorithme CEM

Algorithme CEM

- Étape 0 : solution initiale $(\boldsymbol{\theta}^0, \mathcal{Z}^0)$
- Étape E : calcul des

$$t_{ik}^c = \frac{\pi_k^c \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k^c)}{\sum_\ell p_\ell^c \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_\ell^c)}$$

- Étape C : $\mathcal{Z}^{c+1} = \text{MAP}(t_{ki})$
  Les t_{ki}^c sont remplacées par des 1 ou des 0
- Étape M : maximisation de $L(.; \mathcal{X}, \mathcal{Z}^{m+1})$
  - $\pi_k^{c+1} = \frac{\#z_k^c}{n}$
  - $\boldsymbol{\alpha}_k^{c+1}$: estimation max. de vrais. en utilisant la classe k comme échantillon

▶ Exemple du modèle de mélange gaussien

- $\mu_k^{c+1} = \frac{1}{\#z_k^{c+1}} \sum_{i \in z_k^{c+1}} \mathbf{x}_i$
- $\Sigma_k^{c+1} = \frac{1}{\#z_k^{c+1}} \sum_{i \in z_k^{c+1}} (\mathbf{x}_i - \mu_k^{c+1})(\mathbf{x}_i - \mu_k^{c+1})'$
- CEM : algorithme de classification très général

▶ Paramétrisation de la matrice de variance

$$\Sigma_k = \lambda_k D_k A_k D_k'$$

- $\lambda_k = |\Sigma_k|^{\frac{1}{d}}$ nombre positif : volume de la classe k
- A_k matrice diagonale de déterminant 1 avec des valeurs allant en décroissant : forme de la classe k
- D_k matrice orthogonale des vecteurs propres : orientation de la classe k

Finalement les paramètres du modèle de mélange gaussien sont :

- les centres des classes $\mu_1, \ldots, \mu_g$
- les proportions $\pi_1, \ldots, \pi_g$
- les volumes $\lambda_1, \ldots, \lambda_g$
- les formes $A_1, \ldots, A_g$
- les orientations $D_1, \ldots, D_g$

▶ Variantes du modèle de mélange gaussien

En jouant sur différentes caractéristiques du modèle :

- Formes sphériques (I), diagonales (B) ou quelconques (C)
- Proportions, volumes, formes et directions identiques ou non suivant les classes,

plusieurs variantes peuvent être proposées :

- $[\boldsymbol{\pi}, \lambda I]$: modèle le plus simple
- $[\boldsymbol{\pi}, \lambda_k I]$:
- $[\boldsymbol{\pi}, \lambda B]$
- ...
- $[\boldsymbol{\pi}_k, \lambda_k C_k]$: modèle le plus compliqué (aucune contrainte)

▶ **Résumé**

Modèle	distance	critère	Remarques
$\boldsymbol{\pi}, \lambda I$	$d^2(\mathbf{x}_i, \mu_k)$	trace(W)	CEM = k-means
$\boldsymbol{\pi}, \lambda_k I$	$\frac{1}{\lambda_k} d^2(\mathbf{x}_i, \mu_k) + d \ln(\lambda_k)$		surf. separ. : hypersph.
$\boldsymbol{\pi}, \lambda B$	$d^2_{B^{-1}}(\mathbf{x}_i, \mu_k)$	diag(W)	classification + pond.
$\boldsymbol{\pi}, \Sigma$	$d^2_{\Sigma^{-1}}(\mathbf{x}_i, \mu_k)$	$\lvert W \rvert$	

▶ Utilisation des modèles de mélange en discrimination

- Estimation des paramètres
- Utilisation de cette estimation pour affecter un nouvel individu : classifieur plug-in
- Par exemple, l'analyse discriminante linéaire correspond à un modèle gaussien avec une matrice de variance commune

Apprent. semi supervisé avec le modèle de mélange

- O'Neil (1978) et Ganesalingam et MacLachlan (1978)
- Données : $(\mathcal{X}, \mathcal{Z}_1) = ((\mathbf{x}_1, \mathbf{z}_1), (\mathbf{x}_1, \mathbf{z}_1), \ldots (\mathbf{x}_m, \mathbf{z}_m), \mathbf{x}_{m+1}, \ldots, \mathbf{x}_n)$

$\mathcal{X}$				$\mathcal{Z}$		
3.5	2.3	0.3	4.2	0	0	1
4.2	1.7	2.2	1.1	1	0	0
6.2	1.2	2.2	1.1	0	1	0
2.5	2.3	0.3	4.2	?	?	?
9.2	2.4	2.9	1.3	?	?	?
2.2	1.4	2.9	1.3	?	?	?
7.2	2.9	7.9	1.6	?	?	?

- Problème : estimer $\mathcal{Z}_2 = (\mathbf{z}_{m+1}, \ldots, \mathbf{z}_n)$
- Solution : estimer $\boldsymbol{\theta}$ à partir de $(\mathcal{X}, \mathcal{Z}_1)$ en utilisant EM et appliquer le MAP pour obtenir $\mathcal{Z}_2$

- Données $(\mathcal{X}, \mathcal{Z}_1) = ((\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_m, z_m))$
- Données complètes : $(\mathcal{X}, \mathcal{Z})$
- Données manquantes $\mathcal{Z}_2 = (z_{m+1}, \ldots, z_n)$
- Vraisemblance complétée

$$L(\boldsymbol{\theta}; \mathcal{X}, \mathcal{Z}) = L(\boldsymbol{\theta}; \mathcal{X}_1, \mathcal{Z}_1) + L(\boldsymbol{\theta}; \mathcal{X}_2, \mathcal{Z}_2)$$

- La fonction Q devient

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^c) &= E\left(L(\boldsymbol{\theta}; \mathcal{X}, \mathcal{Z})|\mathcal{X}, \mathcal{Z}_1, \boldsymbol{\theta}^c\right) \\ &= L(\boldsymbol{\theta}; \mathcal{X}_1, \mathcal{Z}_1) + E\left(L(\boldsymbol{\theta}; \mathcal{X}_2, \mathcal{Z}_2)|\mathcal{X}_2, \boldsymbol{\theta}^c\right) \\ &= \sum_{i=1}^{m} \sum_{k} z_{ik} \log \pi_k \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k) \\ &+ \sum_{i=m+1}^{n} \sum_{k} P(z_{ik} = 1|\mathbf{x}_i, \boldsymbol{\theta}^c) \log \pi_k \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k) \end{aligned}$$

- Algorithme EM
  - étape E : calcul de $t_{ik}^c = P(\mathbf{x}_i \in P_k | \mathbf{x}_i, \pi^c, \boldsymbol{\alpha}^c)$ pour $i \in \mathcal{X}_2$

$$t_{ik}^c = \frac{\pi_k^m \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k^c)}{\sum_{\ell=1}^g \pi_\ell^c \varphi_\ell(\mathbf{x}_i; \boldsymbol{\alpha}_\ell^c)}$$

  - étape M : maximisation de la vraisemblance cond. aux t_{ik}^c

- Exemple : calcul de la moyenne pour un mélange gaussien
  - Tous les individus ont une classe connue

$$\mu_k^{c+1} = \frac{\sum_i z_{ik}\mathbf{x}_i}{\sum_i z_{ik}}$$

  - Aucun individu n'a une classe connue

$$\mu_k^{c+1} = \frac{\sum_i t_{ik}^c \mathbf{x}_i}{\sum_i t_k^c(\mathbf{x}_i)}$$

  - Cas général

$$\mu_k^{c+1} = \frac{\sum_{i=1}^n z_{ik}\mathbf{x}_i + \sum_{i=m+1}^n t_{ik}^c \mathbf{x}_i}{\sum_{i=1}^n z_{ik} + \sum_{i=m+1}^n t_k^c(\mathbf{x}_i)}$$

Apprentissage partiellement supervisé avec le modèle de mélange

- Données : $(\mathcal{X}, \mathcal{Z}_1)$

3.5	2.3	0.3	4.2	0	0	1
2.2	1.4	2.9	1.3	0	?	?
4.2	1.7	2.2	1.1	0	0	1
2.5	2.3	0.3	4.2	?	?	?
4.3	2.9	6.2	4.9	?	0	?
9.2	2.4	2.9	1.3	?	?	?
6.2	1.2	2.2	1.1	1	0	0

- Problème : estimer $\mathcal{Z}_2$ = les z_{ik} manquants
- Solution : estimer $\boldsymbol{\theta}$ à partir de $(\mathcal{X}, \mathcal{Z}_1)$ en utilisant EM et appliquer le MAP pour obtenir $\mathcal{Z}_2$

- Données $(\mathcal{X}, \mathcal{Z}_1)$
- Données complètes : $(\mathcal{X}, \mathcal{Z})$
- Données manquantes $\mathcal{Z}_2$
- Vraisemblance complétée

$$L(\boldsymbol{\theta}; \mathcal{X}, \mathcal{Z}) = \sum_i \sum_k z_{ik} \log\{\pi_k \varphi_k(\mathbf{x}_i, \boldsymbol{\alpha}_k)\}$$

- La fonction Q devient

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^c) &= E\left(L(\boldsymbol{\theta}; \mathcal{X}, \mathcal{Z})|\mathcal{X}, \mathcal{Z}_1, \boldsymbol{\theta}^c\right) \\ &= \sum_{i,k \in \mathcal{Z}_1} z_{ik} \log\{\pi_k \varphi_k(\mathbf{x}_i, \boldsymbol{\alpha}_k)\} \\ &+ \sum_{i,k \in \mathcal{Z}_2} P(z_{ik}|\mathbf{x}_i, \boldsymbol{\theta}^c) \log\{\pi_k \varphi_k(\mathbf{x}_i, \boldsymbol{\alpha}_k)\} \end{aligned}$$

- Algorithme EM
  - étape E : calcul de $t_{ik}^c = P(\mathbf{x}_i \in P_k | \mathbf{x}_i, \pi^c, \boldsymbol{\alpha}^c)$ pour i, k vérifiant $z_{ik} \in \mathcal{Z}_2$

$$t_{ik}^c = \frac{\pi_k^m \varphi_k(\mathbf{x}_i; \boldsymbol{\alpha}_k^c)}{\sum_{\ell / z_{i\ell} \neq 0} \pi_\ell^c \varphi_\ell(\mathbf{x}_i; \boldsymbol{\alpha}_\ell^c)}$$

  - étape M : maximisation de la vraisemblance cond. aux t_{ik}^c

Application à la fouille d'images

- Objectif : développer un algorithme d'aide à la consultation d'une base de dossiers médicaux comportant des images
- Données : images médicales (IRM des mains)
- Deux classes :
  - images pertinentes (choisies par l 'utilisateur)
  - images non pertinentes

- Particularités
  - Très faible nombre d'images annotées en début de recherche (petite taille de l'ensemble d'apprentissage)
  - Dimension souvent importante de l'index (nombre important de variables)
  - Définition incrémentale de l'ensemble d'apprentissage
  - Les paramètres du classifieur doivent être remis à jour rapidement après chaque nouveau bouclage de pertinence
- Idée : utilisation des modèles de mélange hiérarchique dans un cadre semi-supervisé

- **Modèles hiérarchiques**
  - MDA : Mixture Discriminant Analysis
  - Hastie et Tibshirani (JRSS, 1994)
  - Généralise l'utilisation du modèle de mélange en discrimination
  - Chaque classe est modélisée par un ou plusieurs composants d'un modèle de mélange

- Conséquence : l' apprentissage semi-supervisé devient partiellement supervisée :

$\mathcal{X}$				$\mathcal{Z}$	
3.5	2.3	0.3	4.2	?	?
2.2	1.4	2.9	1.3	1	0
4.2	1.7	2.2	1.1	0	1

$\mathcal{X}$				$\mathcal{Z}$		
3.5	2.3	0.3	4.2	?	?	?
2.2	1.4	2.9	1.3	?	?	0
4.2	1.7	2.2	1.1	0	0	1

▶ Problème à 2 classes

- Classe des images pertinentes : mélange de 2 lois gaussiennes à matrice de variance diagonale
  - ➤ plusieurs centres d'intérêt possibles
  - ➤ nombre de paramètres raisonnable
- Classe des images non pertinentes : loi uniforme
  - ➤ images considérées comme du bruit
  - ➤ un seul paramètre à estimer

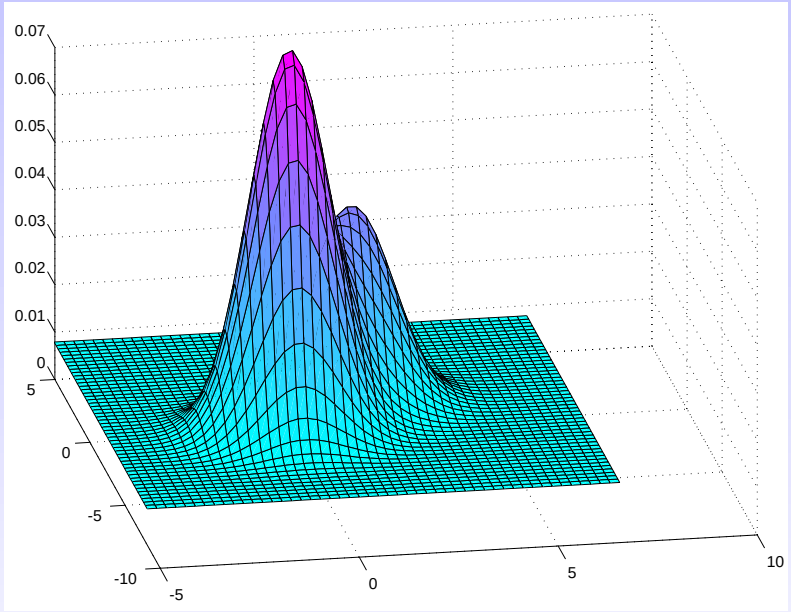
▶ En pratique

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_{k=1,2} \pi_k \varphi_k(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) + \pi_3 . \frac{1}{V}, \quad (1)$$

avec

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \quad (2)$$

$\boldsymbol{\mu}_k$ et $\boldsymbol{\Sigma}_k$ sont les vecteurs moyennes et matrices de covariances des composantes de la classe des images pertinentes. V est le volume du domaine des index



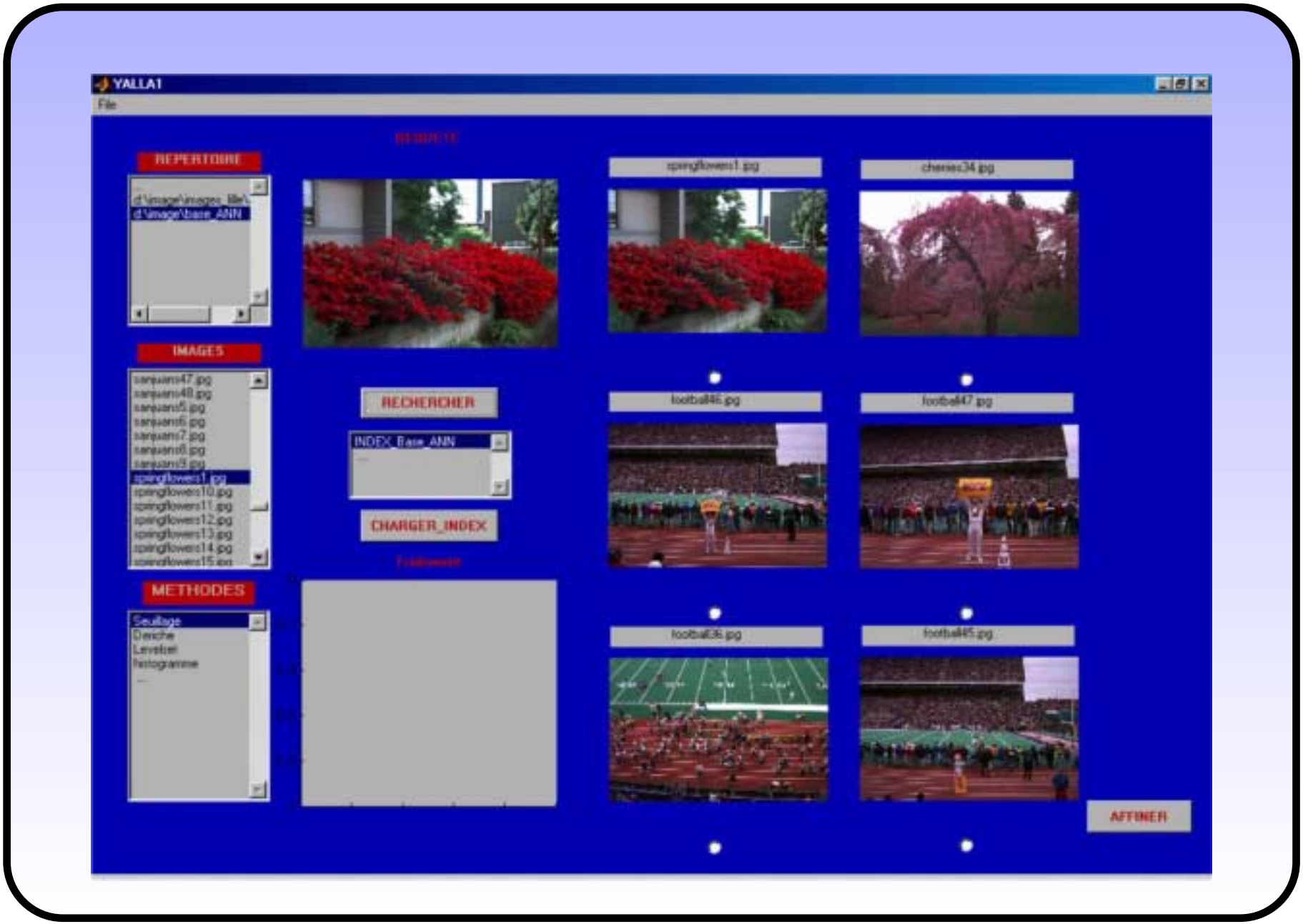


FIG. 3 – requête fleurs itération 1

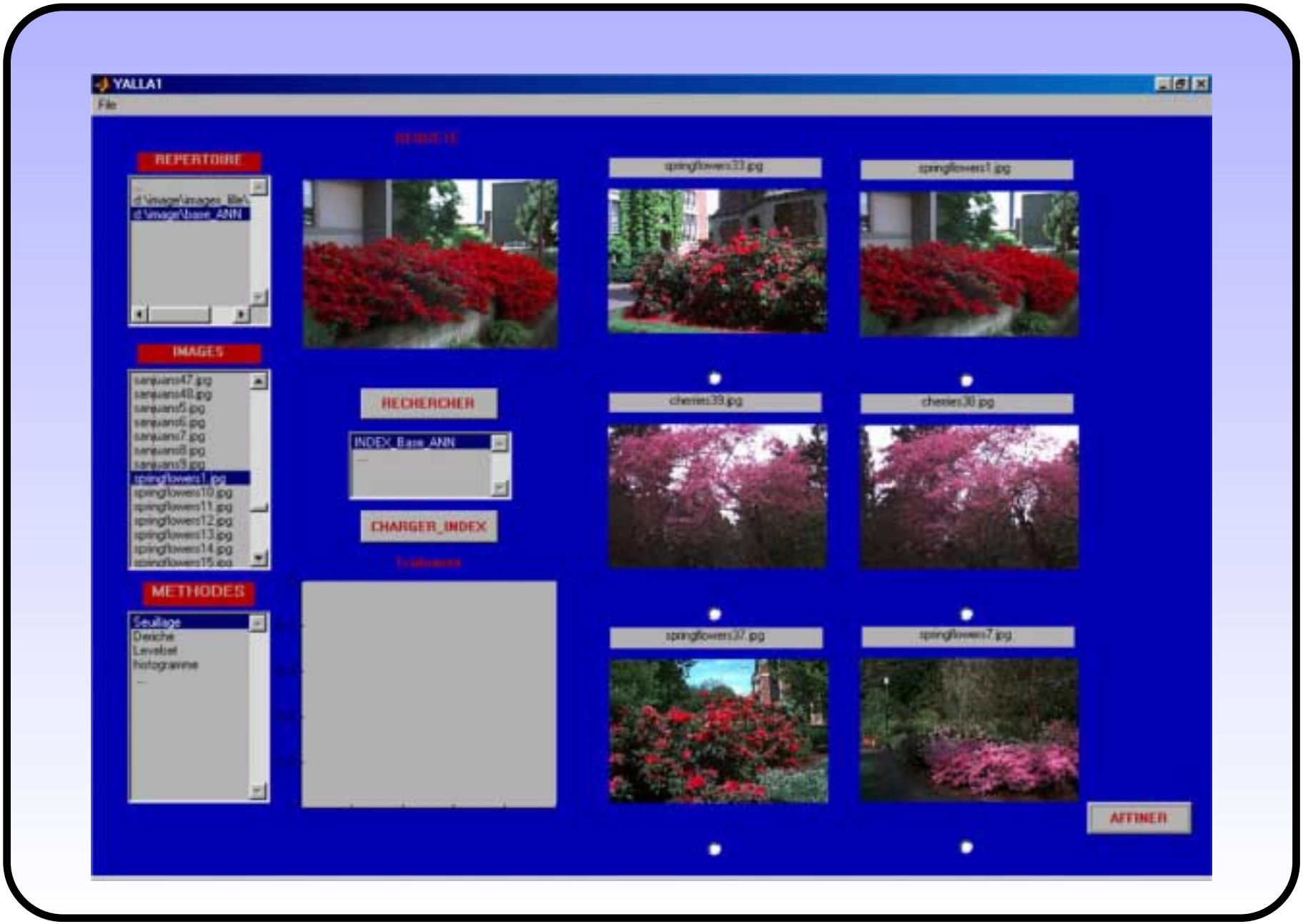


FIG. 4 – requête fleurs itération 3

Remarques conclusives

▶ Modèles de mélange avec des données partiellement supervisées

- Prise en compte simultanément
  - ➤ des observation non étiquetées
  - ➤ des observations étiquetées
  - ➤ des observations partiellement étiquetées
- Approche intermédiaire entre clustering et discrimination
- Idées simples à implémenter
- Peut éviter des problèmes de singularité
- Permet généralement d'accélérer l'algorithme EM
- Solution pour le choix de la situation initiale

- **Extension aux méthodes dérivées de EM**
  - CEM (Cetim)
  - SEM
  - ...
- **Autre approche envisageable (mélange ou clustering)**

Utiliser uniquement les données labellées pour « régler »l'algorithme retenu :

- Nombre de classes
- Situation initiale
- Coefficient de régularisation (par exemple, en spatial)
- ...

- **Problème posé : évaluation de la performance**
  - Existence de critère pour le supervisé : % de bien classé,...
  - Existence de critère pour le non supervisé : vraisemblance
  - Partiellement supervisé : plus compliqué
    - Exemple : classification de textes (Nigam)
    - Modèle de mélange particulier (*Naive Bayes classifier*)
    - Introduction d'une pondération entre les individus labellés et non labellés
    - Réglage de cette pondération par validation croisée sur les individus étiquetés

▶ Extension à un étiquetage incertain

- Aitchison et Begg (1976) : étiquettes erronées
- Extension à des informations de type probabiliste
- Données :

3.5	2.3	0.3	4.2	0	0	1
2.2	1.4	2.9	1.3	0	?	?
4.2	1.7	2.2	1.1	0	0	1
2.5	2.3	0.3	4.2	?	?	?
4.3	2.9	6.2	4.9	0.7	0	0.3
9.2	2.4	2.9	1.3	?	?	?
6.2	1.2	2.2	1.1	1	0	0

- Exemple du Cetim

- **Développement d'une approche similaire avec les arbres de décision**
  - Objectif : disposer d'un arbre de décision
  - Classification supervisée : CART
  - Classification non supervisée : classification descendante hiérarchique
  - Classification partiellement supervisée ?

Références

[1] J.A. Anderson. Multivariate logistic compounds. *Biometrika*, 66 :17–26, 1979.

[2] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT : Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, 1998.

[3] A. Demiriz, K. P. Bennett, and M. J. Embrechts. A genetic algorithm approach for semi-supervised clustering. *Journal of Smart Engineering System Design*, 4 :35–44, 2002.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society*, B 39 :1–38, 1977.

[5] S. Ganesalingam and G. J. McLachlan. The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika*, 65 :658–652, 1978.

[6] J. Larsen, A. Szymkowiak, and L. Hansen. Probabilistic hierarchical clustering with labeled and unlabeled data, 2001.

[7] F. H. C Marriott. Optimization methods of cluster analysis. *Biometrika*, 69 :417–421, 1982.

[8] G. J. McLachlan and K. E. Basford. *Mixture Models, Inference and applications to clustering*. Marcel Dekker, New York, 1988.

[9] G.D. Murray and D.M. Titterington. Estimation problems with data from a mixture. *Applied Statistics*, 27 :325–334, 1978.

[10] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3) :103–134, 2000.

[11] T. J. O'neil. Normal discrimination with unclassified observations. *JASA*, 73 :821–826, 1978.

[12] C. Saint-Jean. *Classification paramétrique robuste partiellement supervisée en reconnaissance des formes*. Thèse, Université de la Rochelle, 2001.

[13] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27 :387–397, 1971.

[14] M. J. Symons. Clustering criteria and multivariate normal mixture. *Biometrics*, 37 :35–43, march 1981.