Une méthode de classement sur la base d'une matrice stochastique

Stéphane Verdun Véronique Cariou El Mostafa Qannari

Unité de Sensométrie et de Chimiométrie
ENITIAA de Nantes

Congrès Chimiométrie 2009

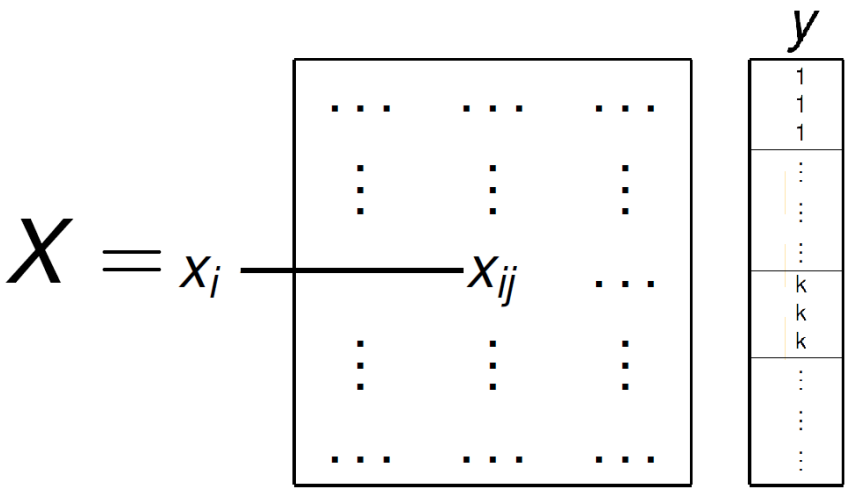
Lignes directrices

1. Introduction
2. Classement basé sur une matrice stochastique
3. Applications
   - Huiles d'olives
   - Données simulées
4. Conclusions

Lignes directrices

1. Introduction
2. Classement basé sur une matrice stochastique
3. Applications
   - Huiles d'olives
   - Données simulées
4. Conclusions

Données et objectifs



← données qualitatives : groupes donnés a priori.

- **Objectif**
  Prédire, à partir de X, les groupes d'appartenance.
- Exemple
  Mesure de la composition selon 8 acides gras sur des huiles provenant de 9 régions d'Italie.
  But : prédire la région de provenance d'une nouvelle huile.

Méthodes usuelles de classement

Méthodes classiques : analyse discriminante linéaire (LDA) et quadratique (QDA).

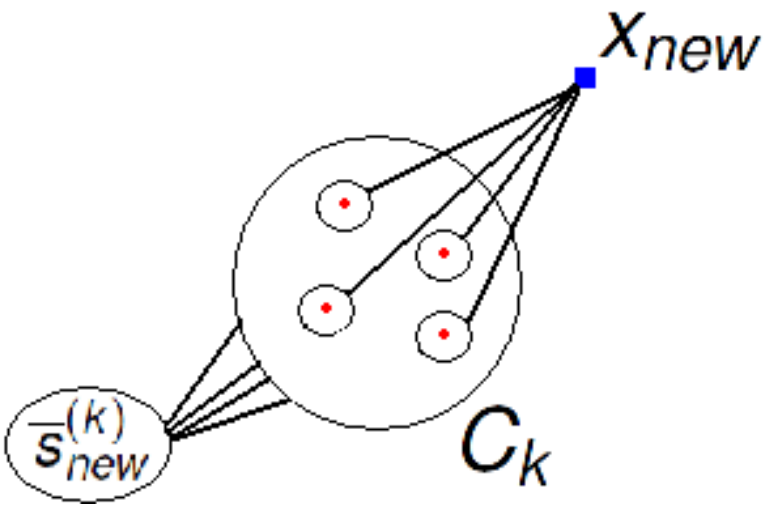
- LDA : $\min_k (x_{new} - \overline{x}_k)' V^{-1} (x_{new} - \overline{x}_k)$
- QDA : $\min_k (x_{new} - \overline{x}_k)' V_k^{-1} (x_{new} - \overline{x}_k)$

avec V matrice de covariance totale et V_k de la classe C_k.

Problèmes

- Quasi-colinéarité des variables : problèmes d'inversion de matrices
- $\overline{x}_k$ peu robuste en présence d'outliers
- Méthodes généralement adaptées pour des distributions multinormales

Réseaux de neurones probabilistes (Specht, 1990)



A l'intérieur de chaque classe C_k, la similarité de x_{new} avec les individus est calculée :

$$s_{new,j} = exp(-\frac{||x_{new} - x_j||^2}{2\sigma^2}).$$

Affectation de x_{new} : $\max_k \bar{s}^{(k)}_{new} = \frac{1}{N_k} \sum_{j=1}^{N_k} s_{new,j}$.

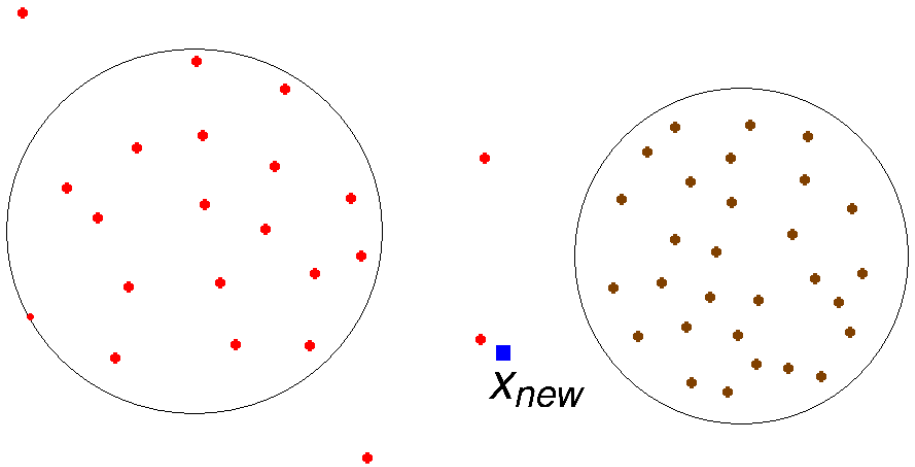
Réseaux de neurones probabilistes

Avantages

- Basée sur la distance euclidienne usuelle donc pas de problèmes d'inversion de matrice
- Ne s'appuie pas sur les barycentres des classes : séparation non linéaire des classes

Inconvénient

- Tous les individus de la classe contribuent de la même manière



Lignes directrices

1. Introduction
2. Classement basé sur une matrice stochastique
3. Applications
   - Huiles d'olives
   - Données simulées
4. Conclusions

Pondération

Première étape

Calcul de la matrice de similarité entre les individus :

$$S = \begin{bmatrix} S^{(1)} & 0 & 0 & 0 \\ 0 & S^{(2)} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & S^{(K)} \end{bmatrix} \text{ avec } S^{(k)} = (s_{ij}^{(k)})_{i,j \in C_k}$$

Par exemple :

- $s_{ij}^{(k)} = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$,
- $s_{ij}^{(k)} =$ Plus proches voisins,
- ...

Pondération

Deuxième étape

Passage de la matrice S à une matrice stochastique P (normalisation) :

$$P = \begin{bmatrix} P^{(1)} & 0 & 0 & 0 \\ 0 & P^{(2)} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & P^{(K)} \end{bmatrix} \text{ avec } P^{(k)} = (p_{ij}^{(k)})_{i,j \in C_k}$$

Pondération

Troisième étape

- Extraire le vecteur de probabilité stationnaire de chaque matrice $P^{(k)}$, i.e. le vecteur $\pi^{(k)}$ tel que :

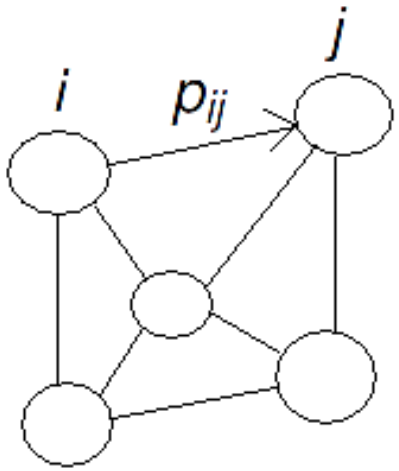
$$\pi^{(k)} P^{(k)} = \pi^{(k)}$$

- Les composantes de $\pi^{(k)}$ sont assignées comme poids aux individus de la classe k
- Affecter un nouvel individu au groupe qui maximise la moyenne pondérée des similarités :

$$\max_k \bar{s}_{new}^{(k)} = \sum_{j=1}^{N_k} \pi_j^{(k)} s_{new,j}$$

Justification / Interprétation

- Les probabilités stationnaires $\pi^{(k)}$ peuvent être interprétées comme une mesure de l'importance des individus
- Théorie des graphes : chaque individu est associé à un sommet d'un graphe
- Marches aléatoires : la matrice P est une matrice de passage d'une marche aléatoire sur les sommets du graphe



Lignes directrices

1 Introduction

2 Classement basé sur une matrice stochastique

3 Applications
- Huiles d'olives
- Données simulées

4 Conclusions

Lignes directrices

1 Introduction

2 Classement basé sur une matrice stochastique

3 Applications
- Huiles d'olives
- Données simulées

4 Conclusions

Présentation des données

- 572 individus (huiles d'olives italiennes).
- 8 variables (compositions selon 8 acides gras).
- 9 classes (régions de provenance).

Objectif :
Prédire la région de provenance d'une nouvelle huile.

Forina, M., et al., *Classification of olive oils from their fatty acid composition.* Food Research and Data Analysis, 1983

Choix du paramètre

- Utilisation de la similarité Gaussienne :
  $s_{ij} = exp(-||x_i - x_j||^2/(2\sigma^2))$
- Choix du paramètre σ par validation sur un échantillon test
- Échantillon de 400 individus : échantillon d'apprentissage sur lequel le modèle est construit
- Échantillon de 172 individus : échantillon de validation sur lequel le taux d'erreurs de classement est calculé

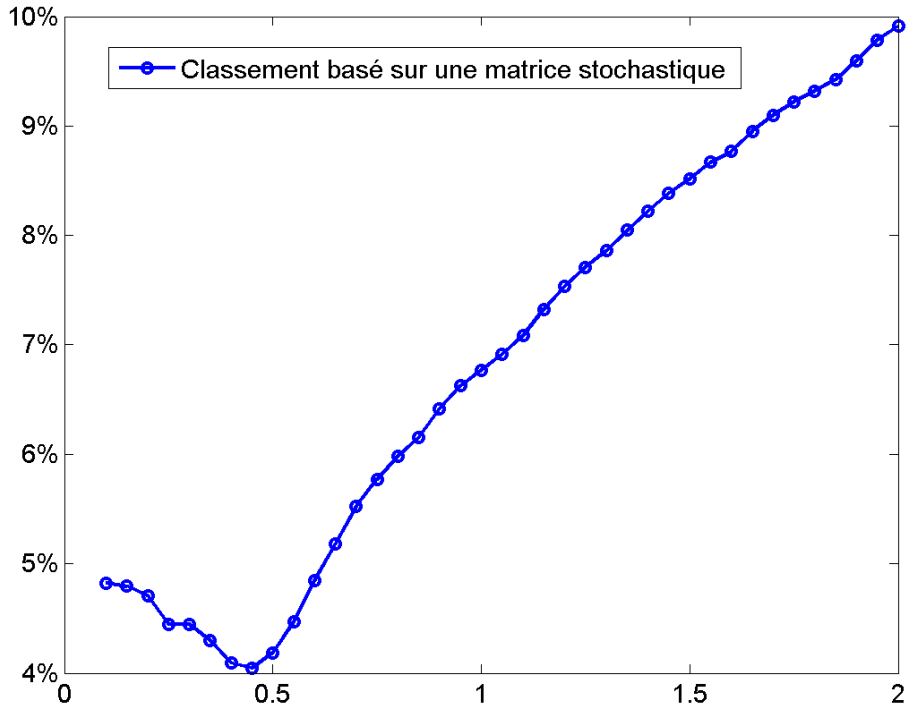


FIGURE: Taux d'erreurs de classement sur l'échantillon de validation en fonction de σ

L'optimum est atteint en $\sigma = 0.45$ avec un taux d'erreurs de classement de 4%.

Comparaison des méthodes

Les individus ont été divisés 100 fois en deux échantillons apprentissage/validation. Plusieurs méthodes ont été appliquées et les taux d'erreurs de classement calculés. Les méthodes comparées sont :

- Analyse discriminante linéaire,
- Analyse discriminante quadratique,
- Réseaux de neurones probabilistes ($\sigma = 0.4$),
- Classement sur la base d'une matrice stochastique.

Comparaison des méthodes

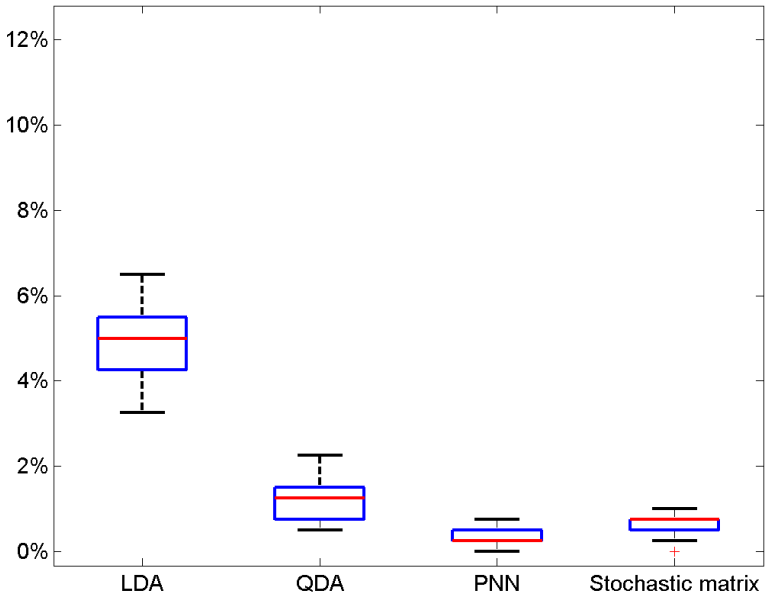


FIGURE: Taux d'erreurs de classement sur l'échantillon d'apprentissage.

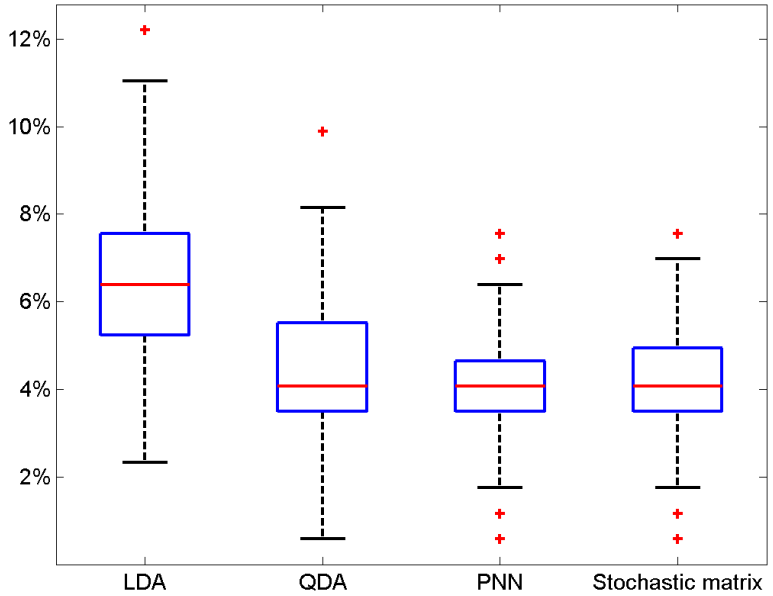


FIGURE: Taux d'erreurs de classement sur l'échantillon de validation.

Lignes directrices

1 Introduction

2 Classement basé sur une matrice stochastique

3 Applications
- Huiles d'olives
- Données simulées

4 Conclusions

Présentation des données

- Données réelles : spectres proches infra-rouge de pommes de différentes variétés
- 1066 individus
- 21 variables (composantes principales issue de l'ACP des spectres)
- 6 groupes
- Données simulées en modifiant l'étiquette de certaines pommes de manière à ce que les classes se chevauchent

Objectif :
Prédire la variété d'une nouvelle pomme.

Résultats

- Paramètre σ estimé sur l'échantillon de validation
- Taux d'erreurs de classement calculés sur 100 échantillons de validation différents

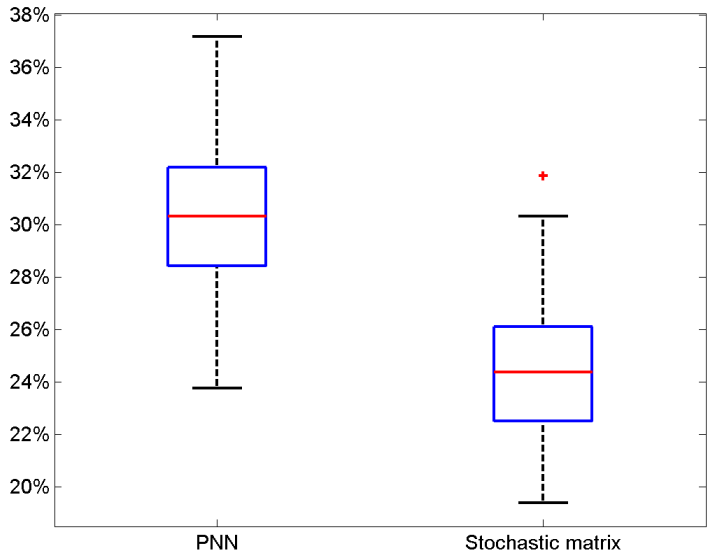


FIGURE: Taux d'erreurs de classement (échantillons de validation)

Lignes directrices

1. Introduction
2. Classement basé sur une matrice stochastique
3. Applications
   - Huiles d'olives
   - Données simulées
4. Conclusions

Conclusions

- Flexible
  - → Choix de la similarité
  - → Résout le problème de sélection des individus avec les poids
- Robuste
  - → Pondération
  - → Prototypes (barycentres) robustes
- Perspectives :
  - → Étudier d'autres similarités (k plus proches voisins, pourcentage de voisins communs...)
  - → Méthodes factorielles basées sur des pondérations selon un schéma ressemblant