Classification

Pr Roch Giorgi

 roch.giorgi@univ-amu.fr

SESSTIM, Faculté de Médecine, Aix-Marseille Université, Marseille, France
http://sesstim.univ-amu.fr

Objectif

- Rechercher une segmentation, partition, des sujets en classes, catégories
- Optimisation d'un critère visant à regrouper les sujets dans des classes
- Homogénéité intra classe
- Hétérogénéité interclasse

Type de Données

- Tableau de distances entre les individus pris 2 à 2
- Variables quantitatives
- Variables qualitatives
- Mélange de variables qualitatives et quantitatives

Méthode

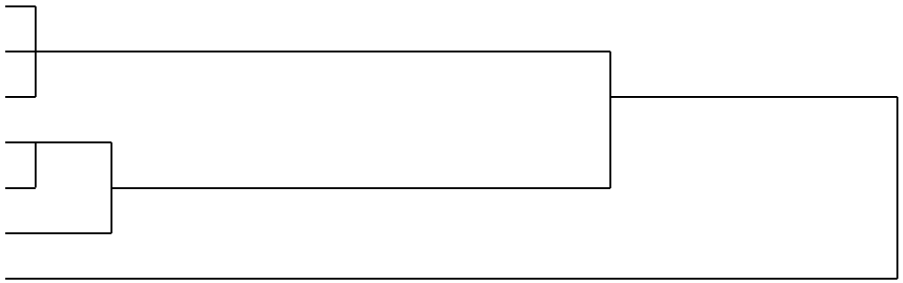
- Le nombre de partitions possibles d'un ensemble de n éléments à une croissance plus qu'exponentielle avec le nombre de sujets
- L'optimisation du critère ne se fait donc pas sur toutes les partitions possibles
- Algorithme itératif convergeant vers une partition optimale

Algorithmes

- Basé sur une mesure d'éloignement entre individus
  - ✓ Dissemblance
  - ✓ Dissimilarité
  - ✓ Distance
- Basé sur l'optimisation du critère d'homogénéité des classes
  - ✓ Traces d'une matrice de variances-covariances
  - ✓ Variances et covariances intra ou interclasses

Classification Ascendante Hiérarchique (1)

- Regroupement des individus par le bas (les 2 les plus proches)
- Agrégation progressive des éléments les plus proches de la partition de l'étape précédente
- Regroupant finalement tous les individus en une seule classe, à la racine
- Construction progressive d'un arbre: dendogramme

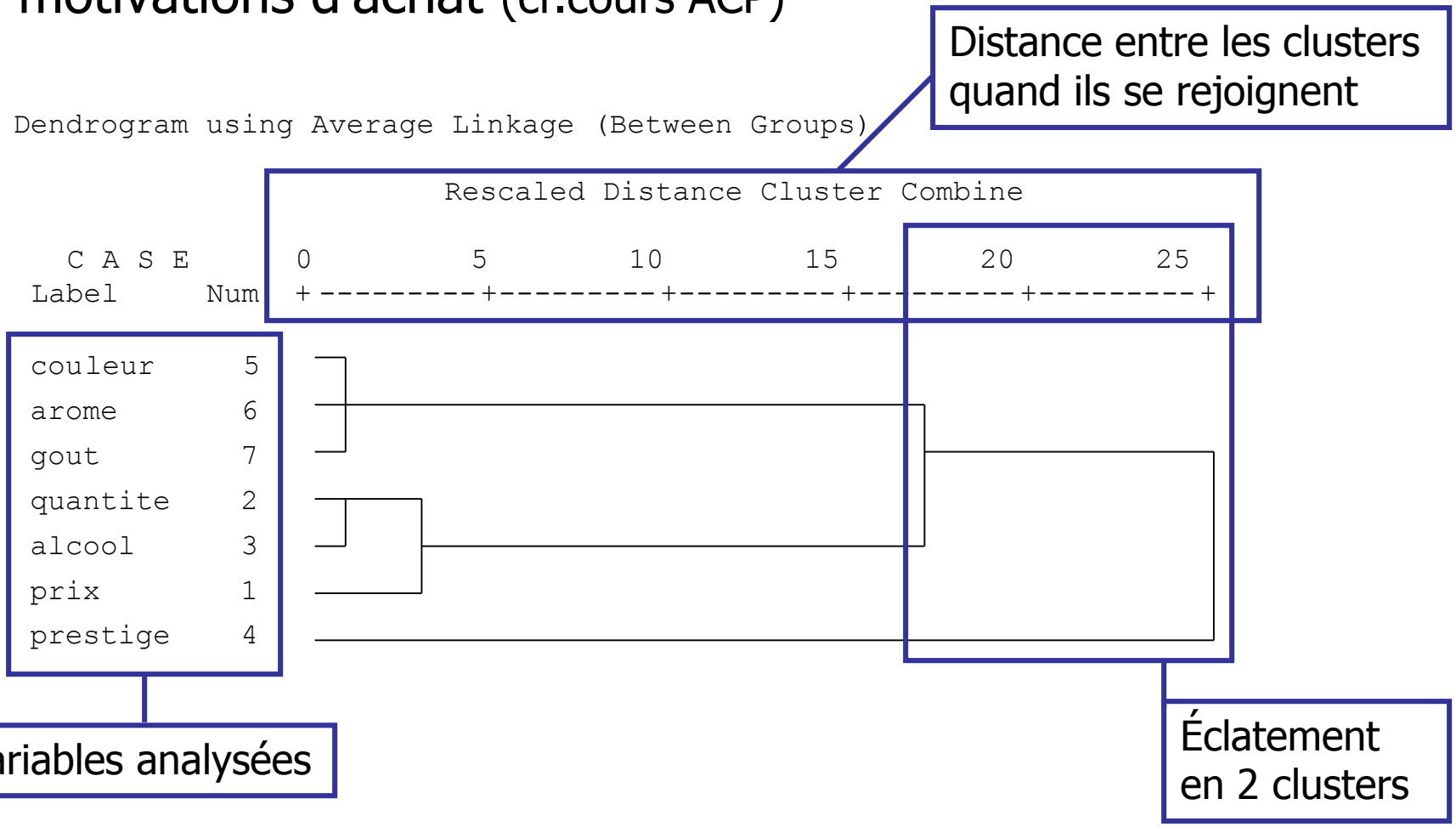


Classification Ascendante Hiérarchique (2)

- Nécessité de calculer à chaque regroupement la dissemblance ou la distance entre un individus et un groupe ou la distance entre 2 groupes
  - ✓ Dissemblance : saut minimum, maximum, moyen
  - ✓ Distance : moyenne entre classes, dans la classe, saut de Ward (minimisation de la variance interclasse)
- La hauteur d'une branche est proportionnelle à l'indice de dissemblance ou de distance entre les 2 objets regroupés
- Le nombre de classes est déterminé *a posteriori*

Exemple

- Enquête sur les consommateurs de bière sur leurs motivations d'achat (cf.cours ACP)

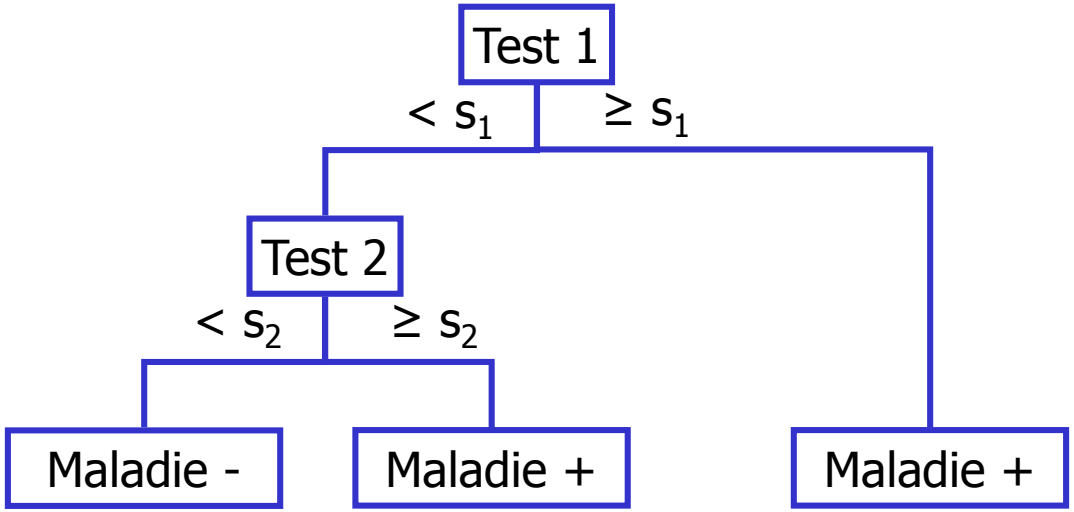


Analyse Discriminante

- Variables qualitatives
- Construire un modèle de prévision de groupe d'affectation (k classes) à partir d'observations
- Établissement de k-1 fonctions discriminantes basées sur les combinaisons linéaires des variables explicatives
- Pour obtenir la meilleur discrimination entre les groupes

Arbres de segmentation (…)

- Arbres de segmentation (Chaid, CRT)
  - ✓ Variable à expliquer qualitative par des variables quantitatives
  - ✓ Détermination
    - De la séquence informationnelle des variables quantitatives
    - Pour des « seuils » optimaux



Phase d'apprentissage
puis
Phase de test

Sources

- Baccini A., Besse A., Déjean S. et coll. Analyse statistique des données d'expression génomique (2005) http://www.lsp.ups-tlse.fr/Besse/pub/Stat_biopuces.pdf
- Besse P., Baccini A. Data mining I - Exploration statistique (2005) http://www.lsp.ups-tlse.fr/Besse/pub/Explo_stat.pdf