ACI « Masse de Données » - 2003

DataHighDim

Analyse exploratoire et discriminante de données en grande dimension

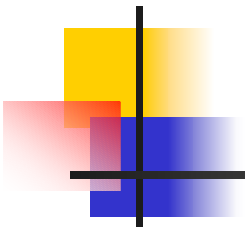
Anne Guérin-Dugué
Laboratoire CLIPS – Grenoble
UJF, CNRS – UMR 5524

CLIPS
Communication Langagière et Interaction Personne-Système
Fédération IMAG
BP 53 - 38041 Grenoble Cedex 9 - France

Partenaires

- Lab. CLIPS-UJF (Coordinateur), Equipe MRIM, Grenoble
- Lab. LIS-INPG, Equipe SIC, Grenoble
- Equipe SELECT-INRIA FUTUR, Paris Orsay

- Lab. DICE-UCL, Equipe Machine Learning, Louvain-la-neuve, Belgique
- Lab LDG, CEA, Bruyères-le-Chatel



Objectifs

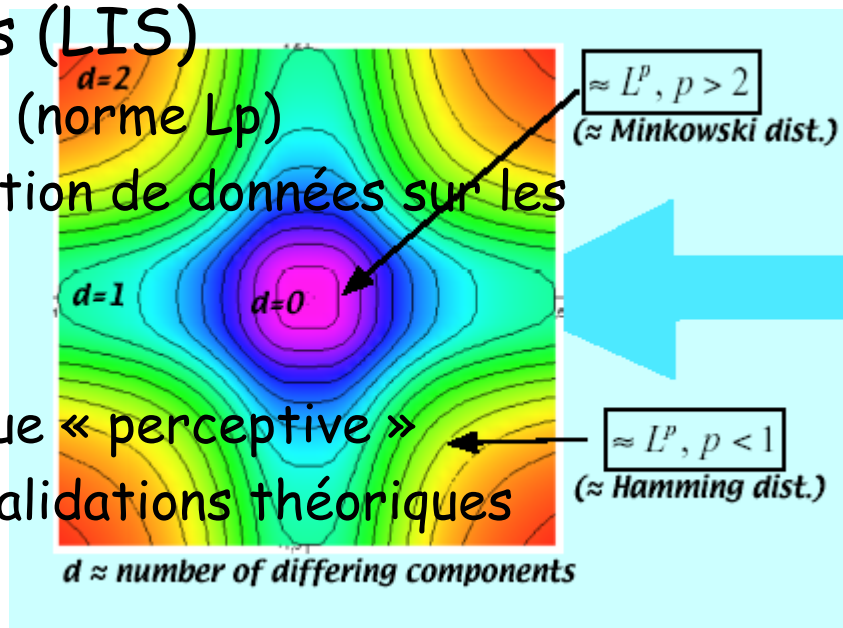
- Analyse Exploratoire et Discriminante de données en grande dimension
  - Méthodes et Outils
- Masse de données, Grande dimension ?
  - Plusieurs dizaines de milliers d'observations
  - Une centaine de variables (dimension)
- Données ?
  - Tableaux « observations x individus »
  - Tableaux croisées de dissimilitudes
  - Données manquantes

Thèmes abordés

- Mesure de distances (LIS, UCL)
- Représentation en basse dimension par projection non linéaire (LIS, SELECT, CEA)
- Analyse discriminante sur tableaux de dissimilitudes (SELECT, CLIPS)
- Applications
  - Données sismiques (données fournies par le CEA)
  - Navigation dans les bases d'images (données fournies par le CLIPS)

Mesures de distance

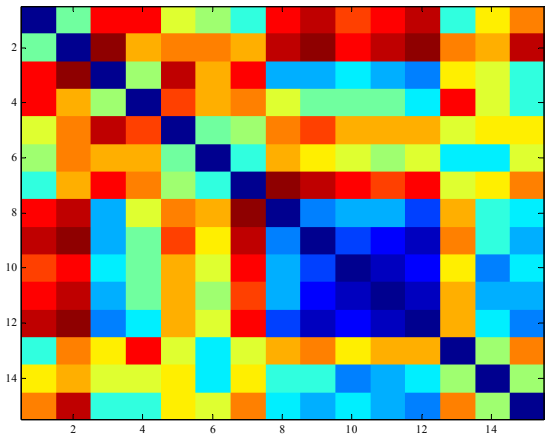
- Concentration de la norme (UCL)
  - Année 1 : Analyse des travaux de [Beyer et al, 1999], Normes Lp (p<1) et Recherche au PPV
  - Année 2, 3 : Conséquence d'un choix de p<1 en pratique (données réelles et simulées)
- Rang d'une matrice de distances (LIS)
  - Année 1 : Analyse en fonction de p (norme Lp)
  - Année 2, 3 : Application à la réduction de données sur les tableaux de distances
- Metrique alternative (LIS)
  - Année 1 : Proposition d'une métrique « perceptive »
  - Année 2,3 : Expérimentations et validations théoriques



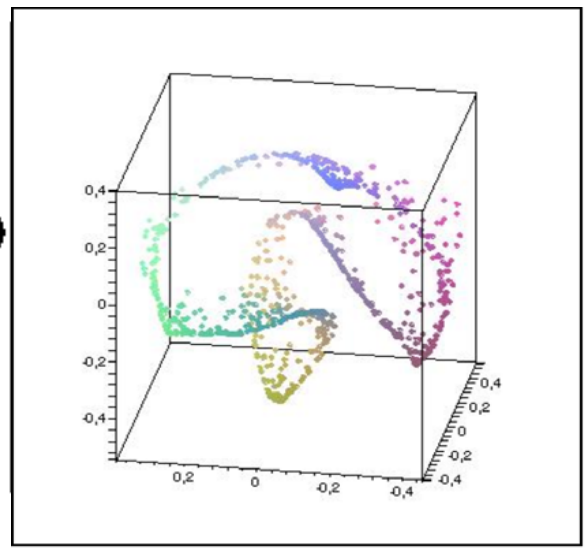
Représentation en basse dimension par projection Non Linéaire (1/2)

CCA : Curvilinear Components Analysis [Demartines, Hérault, 1997]

LIS



CCA

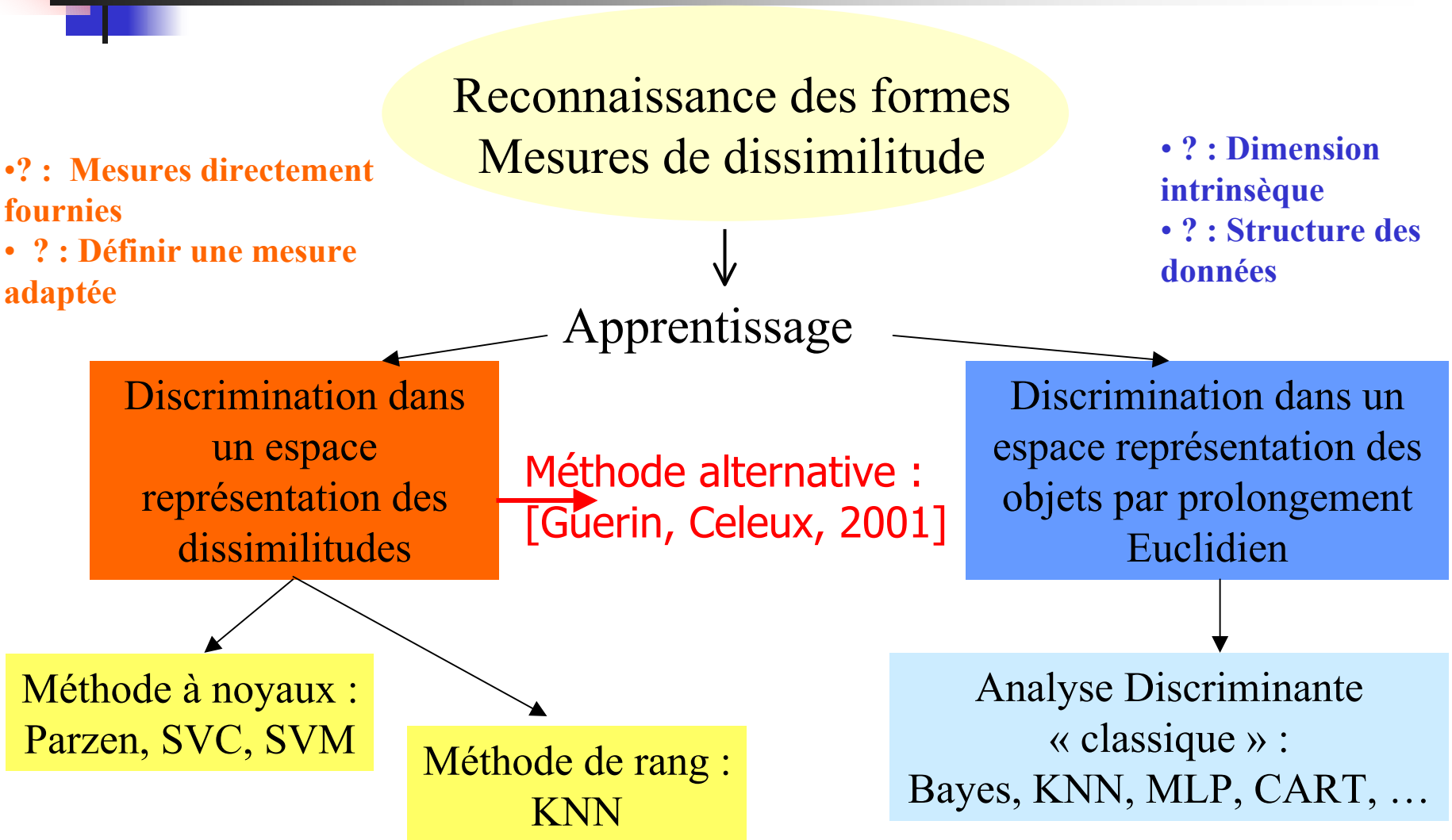


- Année 1
  - Production d'un logiciel d'ACC
  - Expérimentation avec des données de dissimilitudes
- Année 2,3
  - Amélioration de la fonction de coût
  - Expérimentation avec des normes Lp

Représentation en basse dimension par projection Non Linéaire à partir de données de dissimilitudes (2/2)

- Approche probabiliste par MV (SELECT)
  - Année 1 :
    - ✓ Etude des algorithmes (approche Bayesienne), proposés par [Oh, Raftery, 2001] pour la représentation et la discrimination conjointe
  - Année 2, 3 :
    - ✓ Proposition de critère d'information pour le choix de la dimension de représentation en sortie
    - ✓ Proposition d'un modèle, expérimentations

Analyse discriminante sur tableaux de dissimilitudes (1/2)



Analyse discriminante sur tableaux de dissimilitudes (2/2)

- Développement de l'algorithme [Guerin, Celeux, 2001] (CLIPS, SELECT)
  - Année 1 : Extension de la méthode à K (>2) classes
  - Année 2, 3 : Amélioration des procédures d'optimisation des paramètres (apprentissage)
- Approche bayesienne (SELECT)
  - Année 2, 3 : faisabilité pratique pour des grands tableaux, solutions alternatives d'optimisation

Applications

- Données sismiques (fournies par le CEA)
  - Réseaux de capteurs : Classification de l'activité sismique
  - Comparaison avec les approches menées par le CEA (analyse exploratoire et discriminante)
- Dissimilitudes inter « Images » (fournies par le CLIPS)
  - Navigation dans les bases d'images (Représentation de données de dissimilitudes)
  - Catégorisation d'images (Analyse discriminante)

Difficultés, Bilan

- Manque de ressources humaines
  - Pas de post-doc (?), pas de thésard
  - 1 à 2 stagiaires en M2R
- Réunions de travail (3/an)
  - Tous les partenaires
  - 1 réunion avec « MoviStar »
- Demande d'action incitative « Tournesol » franco-belge
  - Réponse fin 2004
- Année 1
  - Redéfinition des objectifs / (financement + RH)
  - Début des travaux sur les 3 thèmes maintenus