le cnam

Classification de données longitudinales

Ndèye Niang
Conservatoire National des Arts et Métiers

Plan

- Introduction
- Rappels sur la classification
- Distances en analyse de données évolutives
- Typologies de trajectoires

(Référence : analyse de données évolutives méthodes et applications Dazy, Le Barzic Saporta, Lavallard Technip 1996)

Introduction

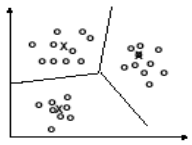
- Données longitudinales multidimensionnelles
- Plusieurs individus - plusieurs variables
- Méthodes « factorielles » permettent visualisation des trajectoires des individus sur les plans factoriels mais interprétation difficile si grand nombre de trajectoires
- Objectif classification = réduction du nombre de trajectoires

Introduction

- Démarche:
  - Application d'une méthode factorielle (STATIS,…)
  - Coordonnées factorielles des individus sur les q axes retenus
  - Application d'une méthode de classification en ayant au préalable défini une distance
  - Les trajectoires des individus de chaque classe sont représentées par la trajectoire du centre de gravité de la classe (réduction)

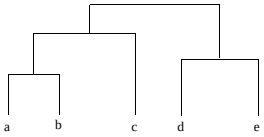
Rappels sur la classification

- **Méthodes de partitionnement :**
- **une partition en un nombre fixe de classes**



- **Méthodes hiérarchiques :**

suite de partitions emboitées



Rappels sur la classification

- **L'objectif des méthodes de classification automatique est la construction d'une partition ou d'une suite de partitions emboîtées d'un ensemble d'objets.**
- **Les classes formées doivent être le plus homogènes possible d'où la nécessite de définir un critère à optimiser.**

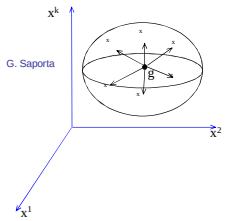
Rappels sur la classification

- **Inertie = « mesure » de la dispersion du nuage de points de R^p espace euclidien**

  I_g par rapport à g centre de gravité

- ***Plus I est faible plus la classe est homogène***

Rappels sur la classification



I = moyenne des carrés des distances à g

$$\sum p_i d^2(i,g)$$

Inertie=variance généralisée $= \sum Var(x_j)$

$$I_W = \sum P_i I_i$$ inertie intra classe

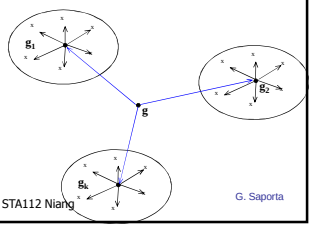
$$I_B = \sum P_i d^2(g_i; g)$$ inertie inter classe

Relation de Huyghens : I = I_W + I_B

partition en k classes

g1 ,..., gk les centres de gravité des classes

I1 , ..., Ik les inerties des k classes



G. Saporta

Rappels sur la classification

- **Critère usuel de classification = chercher la partition qui minimise I_W cad qui maximise I_B**
- **I_W ne permet pas de comparer 2 partitions avec des nombres de classes différents.**

- **Méthode des centres mobiles (Forgy) K-means**

Etape 1

a) configuration initiale : $C_1^{(0)}$, $C_2^{(0)}$ .... , $C_k^{(0)}$

b) chaque individu i est affecté à une classe et une seule $E_i^{(0)}$ de centre $C_i^{(0)}$ telle que : soit minimum en parcourant tous les centres $C_1^{(0)}$, $C_2^{(0)}$ .... , $C_k^{(0)}$

à la fin de cette étape on a k classes $E_1^{(0)}$, $E_2^{(0)}$ .... , $E_k^{(0)}$

Etape 2

a) centres de gravité des classes précedentes : $C_1^{(1)}$, $C_2^{(1)}$ .... , $C_k^{(1)}$

b) chaque individu i est affecté à une classe et une seule $E_i^{(1)}$ de centre $C_i^{(1)}$ telle que : soit minimum en parcourant tous les centres

à la fin de cette étape on a k classes $E_1^{(1)}$, $E_2^{(1)}$ .... , $E_k^{(1)}$

arrêt de la procédure :

- **2 étapes successives ne changent pas les classes**
- **le nombre d'itérations fixé est atteint**
- **la valeur du critère reste inchangée**

Rappels sur la classification

- **Remarque : formes fortes**
- **Problème : la partition finale dépend du nombre de classes et du choix des centres initiaux**
- **Solution : appliquer l'algorithme sur s tirages différents, croiser les s partitions pour obtenir une partition *dite en formes fortes ou regroupements stables***
- **(formes fortes = ensembles d'éléments ayant toujours été regroupés dans la partition finale pour les s passages de l'algorithme).**

Rappels sur la classification

Méthodes hiérarchiques

- **Elles consistent à fournir un ensemble de partitions de E en classes de moins en moins fines par regroupements successifs de parties.**
- **On obtient une hiérarchie représentée par un arbre de classification ou dendrogramme.**
- **On associe au système de classes résultant une échelle de niveau : à chaque partition on associe une valeur numérique représentant le niveau auquel ont lieu les regroupements**
- **Différentes méthodes selon la stratégie de regroupement**

Rappels sur la classification

Méthodes hiérarchiques

Comment mesurer la distance entre une partie et un élément, entre deux parties?

Stratégie d'agrégation:

- saut minimal (single linkage): d(A,B) = inf(a,b)
- diamètre (complete linkage): d(A,B) = sup(a,b)
- Moyenne (average linkage)
- Ward $d_w(A,B) = (p_A p_B/(p_A+p_B))d^2(g_A, g_B)$

Rappels sur la classification

Classification mixte

Les algorithmes classiques sont plus ou moins adaptés à la gestion d'un nombre importants d'objets à classer :

- **partitionnement : ensemble volumineux à faible coût mais la partition dépend des centres initiaux et du nombre de classes.**
- **hiérarchique : non adaptée aux vastes ensembles**

D'où les algorithmes mixtes

- **centres mobiles**
- **classification hiérarchique des groupes obtenus**
- **réaffectation par centres mobiles**

Distances en analyse de données évolutives

- Trois types de distances
  - Distances euclidiennes basées sur les positions
  - Distances euclidiennes basées sur les évolutions des individus
  - Distances compromis positions-évolutions

Distances en analyse de données évolutives

- Données et notations
- Rappel : les coordonnées des individus sont obtenus à l'étape intrastructure par projection des T tableaux sur les axes du compromis
- Tableau de données de la classification = tableau des coordonnées factorielles: q axes retenus * T instants
- Pour chaque individu q*T coordonnées

Distances en analyse de données évolutives

- Voir au tableau

Distances euclidiennes basées sur les positions

- Distance 1: euclidienne classique

$$d_1(i,i')^2 = \sum_{t=1}^{T}\sum_{j=1}^{q}[(c_i^j)^t - (c_{i'}^j)^t]^2$$

- d1(i, i')=0 Ssi égalité de toutes les c(i,j,t)
- On aura une classification selon les positions
- Pas de prise en compte de l'aspect temps

Distances euclidiennes basées sur les évolutions des individus

- On définit l'évolution des individus par:

$$(evol_i^j)^t = (c_i^j)^t - (c_i^j)^{t-1}$$

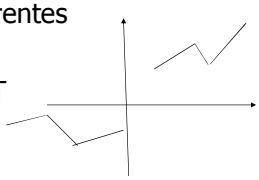
- Et la distance évolution est :

$$d_2(i,i')^2 = \sum_{t=2}^{T}\sum_{j=1}^{q}[((evol_i^j)^t - ((evol_{i'}^j)^t)]^2$$

- d2(i,i') = 0 Ssi les vecteurs de coordonnées sont liés par une translation

Distances euclidiennes basées sur les évolutions des individus

- Plus adaptée aux données temporelles, longitudinales.
- Quelques inconvénients cependant:
- Classe identique car même évolution mais caractéristiques différentes
- Classification de T-1 trajectoires au lieu de T



Distances compromis positions-évolutions

- Distance compromis:
- A. Carlier propose

$$d_3(i,i')^2 = \beta_1 d_1(i,i')^2 + \beta_2 d_2(i,i')^2$$

- Pondération pour « homogénéiser » d1 et d2, choix non explicité
- 2T coordonnées alors que T suffisent

Distances compromis positions-évolutions

Dazy et al proposent:

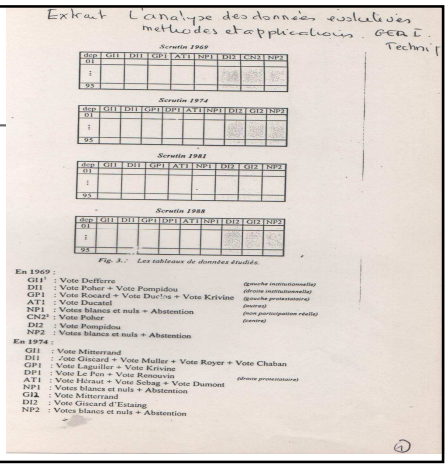
$$d_4(i,i')^2 = \sum_{j=1}^{q} ((c_i^j)^1 - (c_{i'}^j)^1)^2 + d_2(i,i')^2$$

- Premier terme de d1 + d2
  - Point de départ = position à t=1, et T-1 évolutions
- Déséquilibre trop de terme dans d2
- Autre approche:

Distances compromis positions-évolutions

- Principe de l'approche proposée par Dazy et al
- Déterminer les instants t parmi T pour lesquels on utilise la coordonnée-position et ceux pour lesquels on utilise la coordonnée-évolution
- On se ramène à un problème de minimisation d'une fonction différence entre inertie calculée avec les coordonnées-position et inertie calculée avec les coordonnées-évolution

Exemple 1: 1er tour présidentielle


Extrait L'analyse des données évolutives méthodes et applications. GERI. Techni

Scrutin 1969

dep	GI1	DI1	GP1	AT1	NP1	DI2	CN2	NP2
01								
:								
95								

Scrutin 1974

dep	GI1	DI1	GP1	DP1	AT1	NP1	DI2	GI2	NP2
01									
:									
95									

Scrutin 1981

dep	GI1	DI1	GP1	AT1	NP1	DI2	GI2	NP2
01								
:								
95								

Scrutin 1988

dep	GI1	DI1	GP1	DP1	AT1	NP1	DI2	GI2	NP2
01									
:									
95									

Fig. 3.: Les tableaux de données étudiés.

En 1969 :
GI1 : Vote Defferre
DI1 : Vote Poher + Vote Pompidou (gauche institutionnelle) (droite institutionnelle)
GP1 : Vote Rocard + Vote Duclos + Vote Krivine (gauche protestataire)
AT1 : Vote Ducatel (inscrits)
NP1 : Votes blancs et nuls + Abstention (non participation réelle)
CN2 : Vote Poher (inscrits)
DI2 : Vote Pompidou
NP2 : Votes blancs et nuls + Abstention

En 1974 :
GI1 : Vote Mitterrand
DI1 : Vote Giscard + Vote Muller + Vote Royer + Vote Chaban
GP1 : Vote Laguiller + Vote Krivine
DP1 : Vote Le Pen + Vote Renouvin (droite protestataire)
AT1 : Vote Héraut + Vote Sebag + Vote Dumont
NP1 : Votes blancs et nuls + Abstention
GI2 : Vote Mitterrand
DI2 : Vote Giscard d'Estaing
NP2 : Votes blancs et nuls + Abstention

Classification coordonnées position

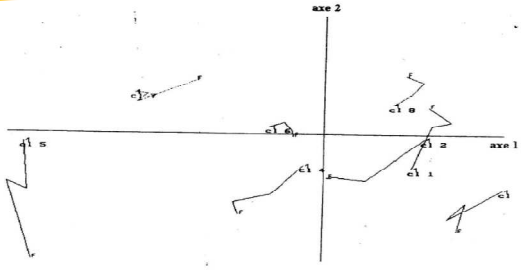
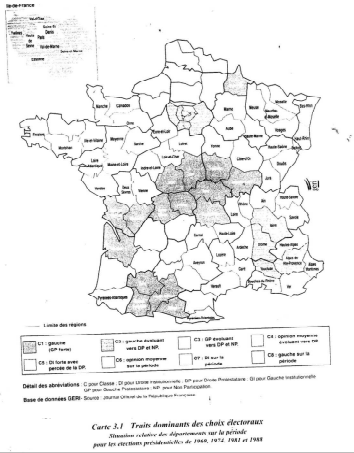


Fig. 3.16 Trajectoires des centres de gravité des classes (classification sur les positions issues de STATIS).


Carte 3.1 Traits dominants des choix électoraux

Application aux données de croissance des jeunes filles

- Après STATIS, on récupère les coordonnées intrastructure des points individus
- On utilise SPAD pour faire la classification basée sur les distances position
- Voir listing document joint.

Remarques, autres cas

- Cas général de la classification de trajectoires: trajectoires de longueur différentes, non synchronisées, …
- Mesures de distances spécifiques: par ex:
  - LCSS (Longest Common Subsequence)
  - DTW (Dynamic Time Warping)
  - Références Vlachos 2003, Aris Anagnostopoulos et Vlachos 2006