

le cnam

Reconnaissance des formes et méthodes neuronales

Chapitre 5 : Classification automatique

Michel Crucianu
<http://idf.pleiad.net/index.php>
<http://cedric.cnam.fr/~crucianm/rfmn.html>

25 septembre 2013 RCP208 1

le cnam

Classification automatique

(cluster analysis, clustering)

- Objectif général : répartir un ensemble donné d'individus en **groupes** (catégories, classes, taxons, *clusters*) de façon à regrouper les individus similaires et à séparer les individus dissimilaires
- Aucune information n'est disponible concernant l'appartenance de certaines données à certaines classes ; le nombre de groupes recherchés peut être connu *a priori* ou non
- Utilisations :
 - ◆ Mettre en évidence une **structure** (simple) dans un ensemble de données
 - ◆ Résumer un grand ensemble de données par les **représentants** des groupes

25 septembre 2013 RCP208 2

Distinctions entre les méthodes

- Par l'objectif :
 - ◆ Partitionnement
 - ◆ Hiérarchie de groupes (→ plusieurs partitionnements disponibles)
- Par la nature des données : numériques, catégorielles, mixtes
- Par la représentation des données :
 - ◆ Riche : représentation vectorielle permettant de définir centre de gravité, densité, intervalles, différentes métriques
 - ◆ Simple : l'unique structure disponible est une métrique
- Par la nature des groupes recherchés : mutuellement exclusifs ou non, nets ou flous
- Par la définition des groupes (critère de regroupement) :
 - ◆ Ensembles **compacts** éloignés entre eux
 - ◆ Ensembles **denses** séparés par des régions moins denses

25 septembre 2013

RCP208

3

Méthode des centres mobiles

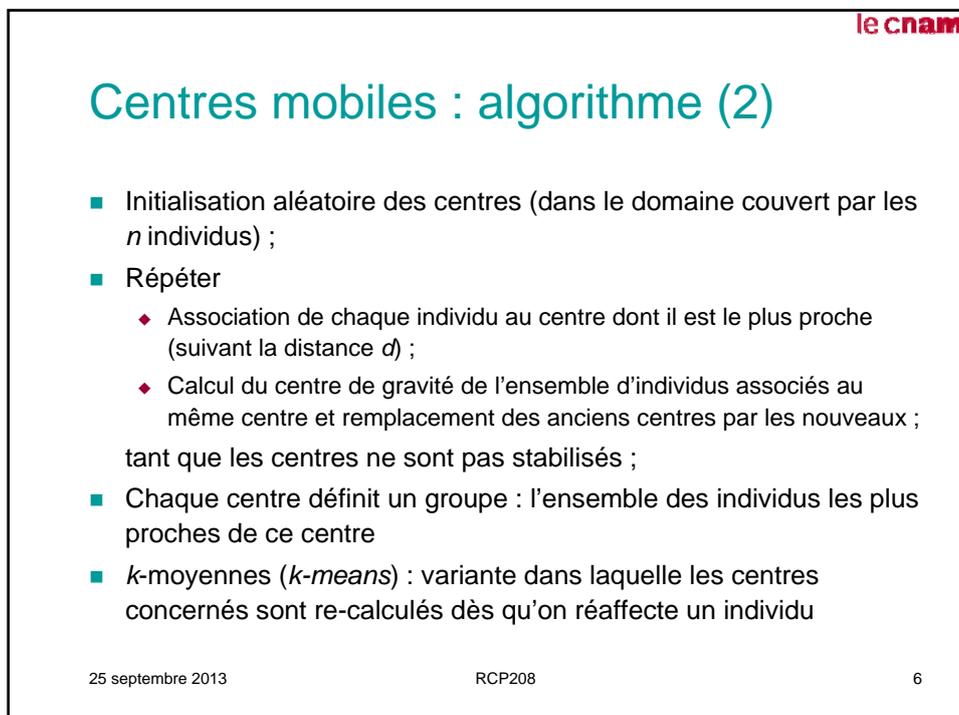
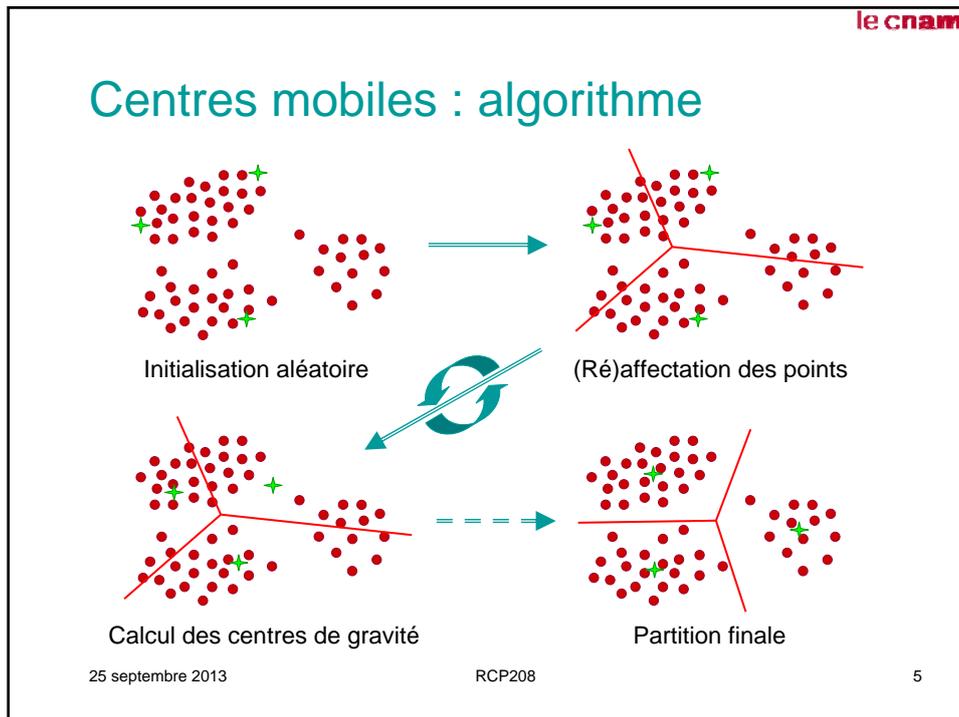
- Objectif : répartir les n individus en k groupes disjoints $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k$ (inconnus *a priori*) en optimisant une fonctionnelle de coût
- Données : ensemble \mathcal{E} de n individus décrits par p variables à valeurs dans \mathbb{R}
- **Distance** dans \mathbb{R}^p (euclidienne usuelle, Mahalanobis, Minkowsky, etc.) qui **définit la nature des (dis)similarités**
- **Fonctionnelle** à minimiser (\mathbf{m}_j : centre de gravité du groupe \mathcal{E}_j) :

$$J(\mathbf{M}) = \sum_{j=1}^k \sum_{\mathbf{x}_i \in \mathcal{E}_j} d^2(\mathbf{x}_i, \mathbf{m}_j) \quad (= \text{somme inerties intra-classe})$$
- Principe de la méthode : à partir d'une initialisation aléatoire de k vecteurs \mathbf{m}_j , alterner entre une étape de (ré) affectation des individus aux groupes (à partir de leur distances aux différents \mathbf{m}_j) et une étape de (re)calcul des centres de gravité

25 septembre 2013

RCP208

4



le cnam

Centres mobiles : illustration

Initialisation Résultat

Démo : www.know-center.tugraz.at/download_extern/WE/clustering/kmeansdemo.inlp

25 septembre 2013 RCP208 7

le cnam

Centres mobiles : convergence

- On peut montrer que la fonctionnelle $J(\mathbf{M})$ diminue (de façon monotone non stricte) lors de chacune des deux étapes du processus itératif :
 - ◆ Association des individus aux centres les plus proches : \mathbf{x}_i passe de \mathbf{m}_p à \mathbf{m}_q si $d^2(\mathbf{x}_i, \mathbf{m}_q) < d^2(\mathbf{x}_i, \mathbf{m}_p)$, impliquant ($\mathbf{m}_{C(j)}$ étant le centre associé à \mathbf{x}_j)

$$\sum_{j \neq i} d^2(\mathbf{x}_j, \mathbf{m}_{C(j)}) + d^2(\mathbf{x}_i, \mathbf{m}_p) > \sum_{j \neq i} d^2(\mathbf{x}_j, \mathbf{m}_{C(j)}) + d^2(\mathbf{x}_i, \mathbf{m}_q)$$
 - ◆ Remplacement des anciens centres par les nouveaux : si $\tilde{\mathbf{m}}_j$ est l'ancien centre du groupe j et \mathbf{m}_j le nouveau, alors $\sum_{\mathbf{x}_i \in \mathcal{E}_j} d^2(\mathbf{x}_i, \tilde{\mathbf{m}}_j) =$

$$\sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{x}_i - \mathbf{m}_j + \mathbf{m}_j - \tilde{\mathbf{m}}_j\|^2 = \sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{x}_i - \mathbf{m}_j\|^2 + \sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{m}_j - \tilde{\mathbf{m}}_j\|^2 +$$

$$+ 2 \sum_{\mathbf{x}_i \in \mathcal{E}_j} {}^T(\mathbf{m}_j - \tilde{\mathbf{m}}_j)(\mathbf{x}_i - \mathbf{m}_j) = \sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{x}_i - \mathbf{m}_j\|^2 +$$

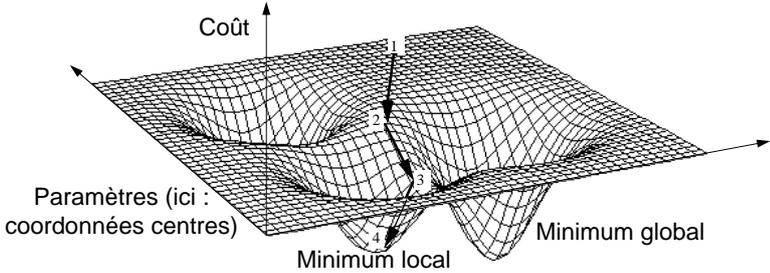
$$\sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{m}_j - \tilde{\mathbf{m}}_j\|^2 + 2 \underbrace{{}^T(\mathbf{m}_j - \tilde{\mathbf{m}}_j) \sum_{\mathbf{x}_i \in \mathcal{E}_j} (\mathbf{x}_i - \mathbf{m}_j)}_{=0} \geq \sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{x}_i - \mathbf{m}_j\|^2$$

25 septembre 2013 RCP208 8

le cnam

Centres mobiles : convergence (2)

- La fonctionnelle $J(\mathbf{M})$ étant positive (donc ayant une borne inférieure), le processus itératif **doit converger**
- ... mais la solution obtenue sera un **minimum local**, dépendant de **l'initialisation** et souvent beaucoup moins bon que le minimum global

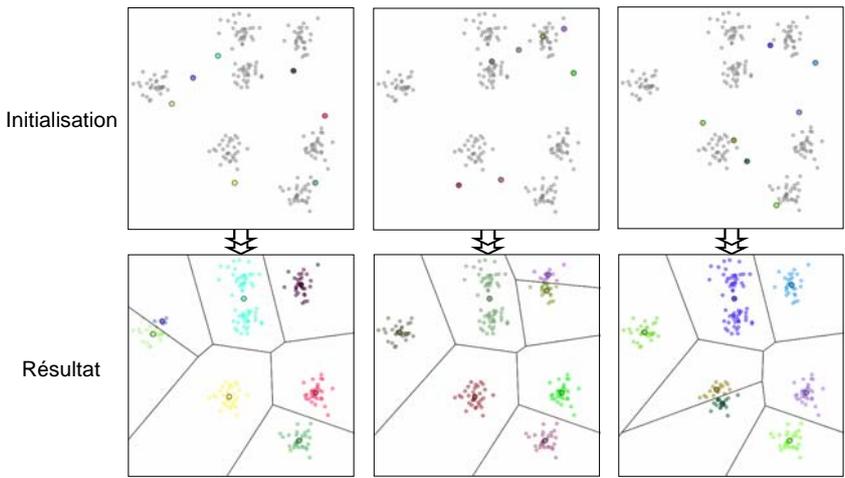


Remarque **importante** : contrairement au cas illustré ci-dessus, $J(\mathbf{M})$ n'est pas différentiable par rapport aux \mathbf{m}_j !

25 septembre 2013
RCP208
9

le cnam

Centres mobiles : illustration (2)



Démonstration : www.know-center.tugraz.at/download_external/WE/clustering/kmeansdemo.jnlp

25 septembre 2013
RCP208
10

Centres mobiles : intérêt et limitations

- Intérêt :
 - ◆ Simplicité...
 - ◆ Paramètre unique : valeur souhaitée pour le nombre de classes
 - ◆ Faible complexité : $O(t \cdot k \cdot n)$ (t étant le nombre d'itérations)
- Limitations → solutions :
 - ◆ Données vectorielles uniquement (pour calculer les moyennes) → limitation levée dans des méthodes dérivées (ex. *k-medoids*)
 - ◆ Classes de forme sphérique (si la distance euclidienne usuelle est employée) → on peut se servir de la distance de *Mahalanobis* (calculée par classe)
 - ◆ Dépendance des conditions initiales (car convergence vers minimum local) → diverses solutions (notamment stochastiques)
 - ◆ Sensibilité au bruit → utilisation d'une fonctionnelle de coût *robuste*, estimation *robuste* des moyennes
 - ◆ Choix *a priori* difficile du nombre de classes → introduction d'un terme de *régularisation* dans la fonctionnelle de coût

25 septembre 2013

RCP208

11

Méthode des *k-medoids*

- Objectif : traiter des *données non vectorielles* (seule une métrique d est connue) en conservant la simplicité des centres mobiles
- *Medoid* d'un groupe = individu le plus « central » du groupe
- Algorithme (une des variantes) :
 1. Choix aléatoire de k individus comme *medoids*
 2. Répéter
 - ◆ Association de chaque individu au *medoid* dont il est le plus proche (suivant la distance d) : $C(\mathbf{x}_i) = \arg \min_j d(\mathbf{x}_i, \mathbf{m}_j)$
 - ◆ Recherche de meilleurs *medoids* et remplacement des anciens par les nouveaux : pour tout j , $\mathbf{m}'_j = \arg \min_{\mathbf{x}_p \in C(\mathbf{m}_j)} \sum_{\mathbf{x}_i \in C(\mathbf{m}_j)} d(\mathbf{x}_i, \mathbf{x}_p)$
 tant que les *medoids* ne sont pas stabilisés ;
 3. Chaque *medoid* définit un groupe : l'ensemble des individus les plus proches de ce *medoid*

25 septembre 2013

RCP208

12

K-medoids : intérêt et limitations

- Avantages :
 - ◆ Fonctionne à partir d'une métrique, donc avec des données non vectorielles
 - ◆ Simplicité conservée
 - ◆ Paramètre unique : valeur souhaitée pour le nombre de classes
 - ◆ Moindre sensibilité au bruit car les distances n'entrent pas au carré dans la fonctionnelle de coût
- Limitations → solutions :
 - ◆ Complexité élevée, $O(t \cdot n^2)$, t étant le nombre d'itérations → emploi de méthodes d'indexation
 - ◆ Classes de forme sphérique (si la distance euclidienne usuelle est employée) → on peut se servir de distances à base de kppv
 - ◆ Choix *a priori* difficile du nombre de classes → introduction d'un terme de régularisation dans la fonctionnelle de coût

25 septembre 2013

RCP208

13

Classification automatique floue

- Objectif : répartir les n individus en k groupes flous $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k$ (inconnus *a priori*) en optimisant une fonctionnelle de coût
- Groupes flous (*fuzzy*) : frontières ne sont pas nettes, un même élément appartiendra à différents degrés à plusieurs groupes
- Données : ensemble \mathcal{E} de n individus décrits par p variables à valeurs dans \mathbb{R} (des extensions à l'utilisation de *medoids* existent)
- Fonctionnelle à minimiser :

$$J(\mathbf{M}) = \sum_{j=1}^k \sum_{i=1}^n u_{ij}^2 d^2(\mathbf{x}_i, \mathbf{m}_j)$$
 où le centre de gravité du groupe \mathcal{E}_j est $\mathbf{m}_j = \frac{\sum_{i=1}^n u_{ij}^2 \mathbf{x}_i}{\sum_{i=1}^n u_{ij}^2}$ et

$$u_{ij} = \frac{1/d^2(\mathbf{x}_i, \mathbf{m}_j)}{\sum_{q=1}^k 1/d^2(\mathbf{x}_i, \mathbf{m}_q)}$$
 est le degré d'appartenance de \mathbf{x}_i au groupe \mathcal{E}_j , avec la contrainte de normalisation $\sum_{j=1}^k u_{ij} = 1$

25 septembre 2013

RCP208

14

C-moyennes floues : algorithme

- Algorithme de classification automatique floue d'individus situés dans \mathbb{R}^p ; \mathbf{m}_j sont les centres de gravité des groupes
- Algorithme (obtenu par minimisation sous contraintes du coût) :
 1. Initialisation aléatoire des centres (dans le domaine couvert par les n individus) ;
 2. Répéter
 - ◆ Calcul du degré d'appartenance de chaque individu à chaque groupe ;
 - ◆ Calcul du centre de gravité de l'ensemble d'individus du même groupe et remplacement des anciens centres par les nouveaux ;
 tant que les centres ne sont pas stabilisés ;
 3. Chaque groupe contient soit tous les individus mais avec les degrés d'appartenance obtenus (groupes flous), soit l'ensemble des individus les plus proches d'un centre (groupes nets, *crisp*)

Agglomération compétitive

(*competitive agglomeration, CA*)

- Objectif : introduire une composante de **régularisation** dans la fonction de coût, permettant à l'utilisateur de **ne pas fixer a priori le nombre de groupes**
- Fonctionnelle à minimiser :

$$J(\mathbf{M}) = \sum_{j=1}^k \sum_{i=1}^n u_{ij}^2 d^2(\mathbf{x}_i, \mathbf{m}_j) - \beta \sum_{j=1}^k \left(\sum_{i=1}^n u_{ij} \right)^2$$

$\beta > 0$ étant un paramètre qui décroît de façon exponentielle au fil des itérations (engendre une « **compétition** » entre les groupes)
- Les degrés d'appartenance sont actualisés suivant

$$u_{ij} = \underbrace{\frac{1/d^2(\mathbf{x}_i, \mathbf{m}_j)}{\sum_{q=1}^k 1/d^2(\mathbf{x}_i, \mathbf{m}_q)}}_{u_{ij}^{\text{FCM}}} + \underbrace{\frac{\beta}{d^2(\mathbf{x}_i, \mathbf{m}_j)}}_{u_{ij}^{\text{Biais}}} \left[N_j - \frac{\sum_{q=1}^k N_q / d^2(\mathbf{x}_i, \mathbf{m}_q)}{\sum_{q=1}^k 1/d^2(\mathbf{x}_i, \mathbf{m}_q)} \right]$$

Agglomération compétitive (2)

- Les **effectifs flous** des groupes, N_j , sont donnés par $N_j = \sum_{i=1}^n u_{ij}$ (avec $\sum_{j=1}^k N_j = n$) et l'évolution de β au fil des itérations t est

$$\beta(t) = \eta_0 e^{-t/\tau} \cdot \frac{\sum_{j=1}^k \sum_{i=1}^n u_{ij}^2 d^2(\mathbf{x}_i, \mathbf{m}_j)}{\sum_{j=1}^k \left(\sum_{i=1}^n u_{ij} \right)^2}$$

- Suite à la compétition, les groupes dont l'effectif flou devient inférieur à un seuil (en général choisi très bas) sont éliminés, ainsi que leurs centres
- Le paramètre « **nombre de groupes** (k) », difficile à choisir, est donc remplacé par la **valeur initiale** et la **vitesse de décroissance** de β , auxquelles l'algorithme est moins sensible (les mêmes résultats sont obtenus pour une plage assez large de valeurs), plus le **seuil d'élimination**, défini à partir de la « résolution » souhaitée

25 septembre 2013

RCP208

17

CA : algorithme

- Algorithme (obtenu par minimisation sous contraintes du coût) :
 - Initialisation aléatoire des centres (dans le domaine couvert par les n individus) ;
 - Répéter
 - Calcul de β pour l'itération courante
 - Calcul du degré d'appartenance de chaque individu à chaque groupe ;
 - Calcul du centre de gravité de l'ensemble d'individus du même groupe et remplacement des anciens centres par les nouveaux ;
 - Calcul de l'effectif flou de chaque groupe et élimination des groupes dont l'effectif est inférieur à un seuil (et des centres correspondants) ;
 tant que les centres ne sont pas stabilisés ;
 - Chaque groupe contient soit tous les individus mais avec les degrés d'appartenance obtenus (groupes flous), soit l'ensemble des individus les plus proches d'un centre (groupes nets, *crisp*)

25 septembre 2013

RCP208

18

Classification floue : intérêt, limitations

- C-moyennes floues :
 - ◆ Avantages : conserve la complexité des centres mobiles, tout en étant moins sensible au bruit grâce à la pondération par les degrés d'appartenance
 - ◆ Limitations : comme pour centres mobiles (moins sensible au bruit)
- Agglomération compétitive :
 - ◆ Avantages :
 - Moins sensible à l'initialisation : l'utilisation de nombreux centres assure leur distribution plus homogène dans les données
 - Ne demande pas à l'utilisateur de fixer *a priori* le « bon » nombre de groupes (dans la pratique, l'algorithme fonctionne bien quand le nombre de groupes est initialisé à une valeur élevée)
 - ◆ Limitations : l'automatisation de la classification reste incomplète...
- Ces méthodes peuvent être étendues à l'utilisation de *medoids*

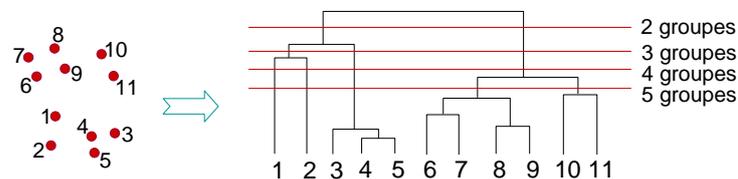
25 septembre 2013

RCP208

19

Classification hiérarchique

- Objectif : obtenir une **hiérarchie de groupes**, qui offre une structure plus riche qu'un simple partitionnement (et permet d'obtenir plusieurs partitionnements à différents niveaux de granularité)



- Méthodes :
 - ◆ Ascendantes : procèdent par agglomération (agrégation) d'individus et de groupes ;
 - ◆ Descendantes : procèdent par partitionnement de l'ensemble initial, puis des groupes

25 septembre 2013

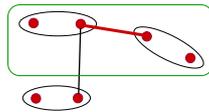
RCP208

20

Classification ascendante hiérarchique

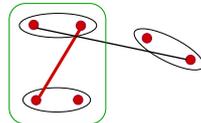
- Procède par agrégation, à chaque étape, des 2 groupes (chaque individu étant également considéré comme un groupe) qui sont les plus proches (les moins dissimilaires)
- Mesures de dissimilarité entre 2 groupes (indices d'agrégation) :

- ◆ Indice du lien minimum (*single linkage*) :



$$\delta(h_p, h_q) = \min_{\substack{\mathbf{x}_i \in h_p \\ \mathbf{x}_j \in h_q}} d(\mathbf{x}_i, \mathbf{x}_j)$$

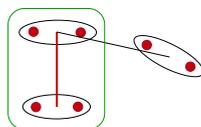
- ◆ Indice du lien maximum (*complete linkage*) :



$$\delta(h_p, h_q) = \max_{\substack{\mathbf{x}_i \in h_p \\ \mathbf{x}_j \in h_q}} d(\mathbf{x}_i, \mathbf{x}_j)$$

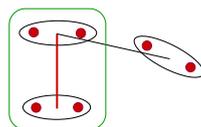
Classification ascendante hiérarchique

- ◆ Indice du lien moyen (*average linkage*) :



$$\delta(h_p, h_q) = \frac{1}{|h_p| |h_q|} \sum_{\substack{\mathbf{x}_i \in h_p \\ \mathbf{x}_j \in h_q}} d(\mathbf{x}_i, \mathbf{x}_j)$$

- ◆ Indice de Ward (augmentation de l'inertie de la réunion des groupes) :



$$\delta(h_p, h_q) = \frac{|h_p| |h_q|}{|h_p| + |h_q|} d^2(\mathbf{m}_p, \mathbf{m}_q)$$

$\mathbf{m}_p, \mathbf{m}_q$: centres de gravité des groupes

Attention : indice de Ward seulement pour **données vectorielles** !

CAH : algorithme

- Remarque : certains indices d'agrégation exigent des données vectorielles, pour d'autres un espace métrique suffit
- Choix nécessaires : mesure de distance entre individus, mesure de dissimilarité entre les groupes (indice d'agrégation)
- Algorithme :
 1. Initialisation : nombre de groupes au départ (N_g) = nombre d'individus (n) ;
 2. Répéter
 - ◆ Calcul de l'indice d'agrégation entre toutes les paires de N_g groupes issues de l'itération précédente
 - ◆ Regroupement des 2 groupes les plus proches (valeur la plus petite de l'indice d'agrégation); les autres groupes restent inchangés ; $N_g := N_g - 1$
 tant que $N_g > 1$;

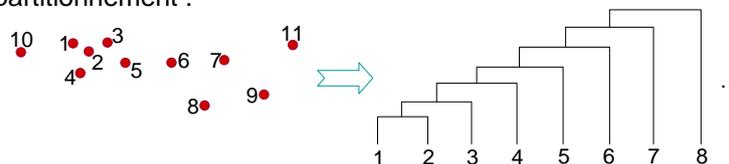
25 septembre 2013

RCP208

23

CAH : effet des différents indices

- Indice du lien minimum : ne tient compte ni de la taille des groupes, ni de leur compacité ; permet de s'approcher d'un critère de regroupement basé sur la densité, mais peut facilement créer des arbres en escalier, déséquilibrés et peu exploitables pour le partitionnement :



- Indice du lien maximum : tient compte de la compacité des groupes, mais pas de leur effectif ; arbres plus équilibrés
- Indice de Ward : tient compte à la fois de l'effectif et de la compacité des groupes résultants ; arbres équilibrés

25 septembre 2013

RCP208

24

CAH : intérêt et limitations

- Avantages :
 - ◆ Satisfait avec une métrique, donc avec des données non vectorielles
 - ◆ Fournit un arbre hiérarchique qui permet d'obtenir **plusieurs partitionnements**, à différents niveaux de granularité
 - ◆ Déterministe (sauf en présence de cas d'égalité entre valeurs de l'indice d'agrégation)
- Limitations → solutions :
 - ◆ Complexité élevée, $O(n^2)$
 - ◆ Choix difficile de l'indice d'agrégation pour un problème donné
 - on tient compte de la nature des groupes recherchés (exemple : groupes compacts → Ward ; groupes denses → lien minimum)
 - examen des résultats obtenus à l'aide de différents indices
 - ◆ Résultats peu exploitables pour certains indices dans certains cas (exemple : effet escalier parfois pour indice du lien minimum)

25 septembre 2013

RCP208

25

Validité de la classification

- La plupart des algorithmes de classification automatique **convergent** vers un résultat **quelles que soient les données**, mais **quelle est la validité de ce résultat** ?
- Principales questions que l'on peut se poser :
 1. Y a-t-il réellement des regroupements « naturels » dans les données ?
 2. Les groupes identifiés sont-ils en accord avec nos (éventuelles) connaissances *a priori* du problème (validation **externe**) ? Remarque : ces connaissances peuvent être confirmatoires et non prescriptives (pas directement exploitables dans la fonctionnelle de coût).
 3. Les groupes identifiés sont-ils bien « ajustés » aux données (validation **interne**) ? Nombreux indices proposés : statistique modifiée de Hubert (alignement entre distance et partition), indice Davies-Bouldin (rapport des inerties), indice de Dunn (rapport diamètre/distance), etc.
 4. Les résultats de la méthode A sont-ils meilleurs que les résultats de la méthode B (validation **relative**) ? En général, application des indices définis pour la validation interne. Possibilités de sélection de modèle.

25 septembre 2013

RCP208

26

Classification semi-supervisée

1. Similarités issues des données (ex. pour des images : descripteurs visuels, mots-clés, relations spatiales...)
 - ◆ Directement calculables là où les données correspondantes sont disponibles
 - ◆ Nombreuses et bon marché
 2. Similarités fournies par les utilisateurs
 - ◆ Explicitement ou à travers des corrélations de recherche, ...
 - ◆ Étiquettes de classe, contraintes, ...
 - ◆ Caractère approximatif, domaine partiel de définition
 - ◆ Plus rares et coûteuses
- **Combiner ces sources** d'information

25 septembre 2013

RCP208

27

Classification semi-supervisée (2)

- La supervision (information concernant la cible : étiquettes de classe, contraintes, ...) n'est disponible que pour une (faible) **partie** des données
 - **apprentissage semi-supervisé**
 - ◆ Question : les différentes sources sont-elles cohérentes ?
 - Coût **élevé** de
 - ◆ l'acquisition de la supervision (exige en général de l'interaction)
 - ◆ l'utilisation des données (complexité algorithmique)
- **apprentissage actif** : sélection par l'algorithme des données pour lesquelles la supervision est demandée
- Amélioration maximale des résultats avec un coût minimal
 - ◆ Question : dans quelles conditions cela **dégrade** les performances ?

25 septembre 2013

RCP208

28

le cnam

Exemple : similarités plus contraintes

- Situation visée (catégorisation de bases d'images) :
 - ◆ Grande base inconnue (ou peu connue)
 - ◆ Classification simple : résultats médiocres
 - Supervision nécessaire !
 - ◆ Classes d'images inconnues *a priori*
 - L'utilisateur ne peut pas donner d'étiquettes mais est capable de dire si 2 images devraient être dans une même classe (contrainte *must-link*) ou dans des classes différentes (contrainte *cannot-link*)
 - Étant donnée la taille de la base, la quantité de supervision (contraintes) doit être minimale



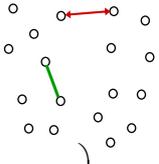
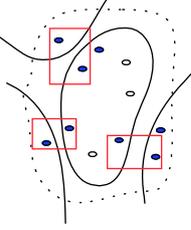

25 septembre 2013
RCP208
29

le cnam

AFCC : principe de la méthode

- Aspect **semi-supervisé**
 - ◆ Combiner deux sources d'information :
 1. Similarités entre descripteurs d'images
 2. Contraintes binaires disponibles
 - Nouvelle fonction à minimiser (basée sur CA) :

$$J(\mathbf{M}) = J_{CA}(\mathbf{M}) + \alpha \left(\sum_{(x_i, x_j) \in \mathcal{M}} \sum_{p=1}^k \sum_{l=1, l \neq p}^k u_{ip} u_{jl} + \sum_{(x_i, x_j) \in \mathcal{C}} \sum_{p=1}^k u_{ip} u_{jp} \right)$$
- Aspect **actif**
 - ◆ Minimiser le nombre de contraintes nécessaires \Leftarrow maximiser le transfert d'information utilisateur \rightarrow système
 - 2 critères de sélection complémentaires :
 1. Contraintes **informatives** : images ambiguës des groupes les moins bien définis
 2. **Faible redondance** entre les contraintes

25 septembre 2013
RCP208
30

AFCC : résultats illustratifs

■ Base *Arabidopsis*



Class 1: 22 plants Class 2: 28 plants



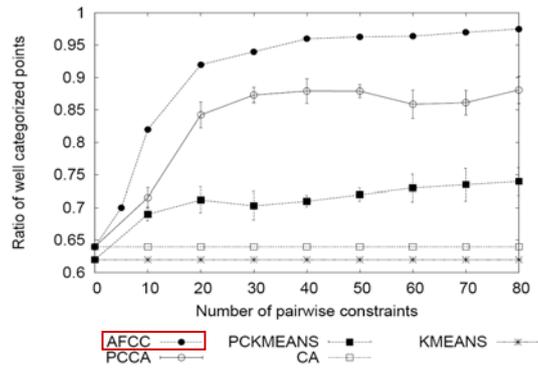
Class 3: 44 plants Class 4: 13 plants



Class 5: 18 plants Class 6: 32 plants



Class 7: 20 plants Class 8: 10 plants



Images fournies par NASC (<http://arabidopsis.info>), vérité terrain INRA (<http://www.inra.fr>)

25 septembre 2013

RCP208

31