Ecole d'ingénieurs du Canton de Vaud

INFORMATIQUE ORIENTATION LOGICIELS

CLASSIFICATION AUTOMATIQUE

Prof. É. D. Taillard

CLASSIFICATION AUTOMATIQUE

But :

Étant donné un ensemble d'éléments, caractérisés par un certain nombre de mesures, on cherche à trouver des sous-ensembles qui soient *homogènes* et *bien séparés*. Les éléments appartenant à un sous-ensemble doivent donc se ressembler et deux éléments appartenant à des sous-ensembles distincts doivent être différents.

Types de classification :

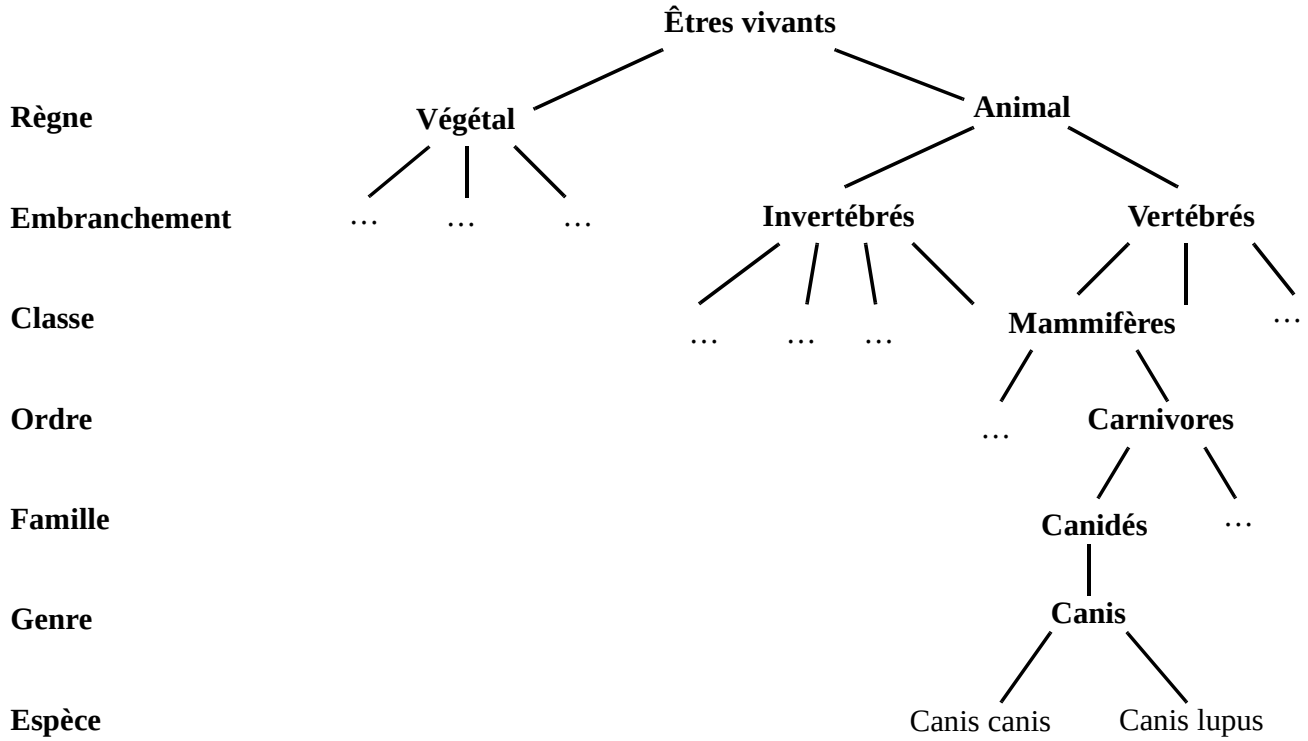
Classification hiérarchique

Les groupes que l'on crée sont emboîtés les uns dans les autres de façon hiérarchique. Exemple : classification des êtres vivants.

Classification non hiérarchique

En classification non hiérarchique, on ne cherche pas à structurer de manière hiérarchique les sous-ensembles entre-eux.

CLASSIFICATION HIÉRARCHIQUE



Méthodes divisives ou agglomératives

En classification *hiérarchique*, on peut procéder par une méthode soit *descendante*, soit *ascendante*.

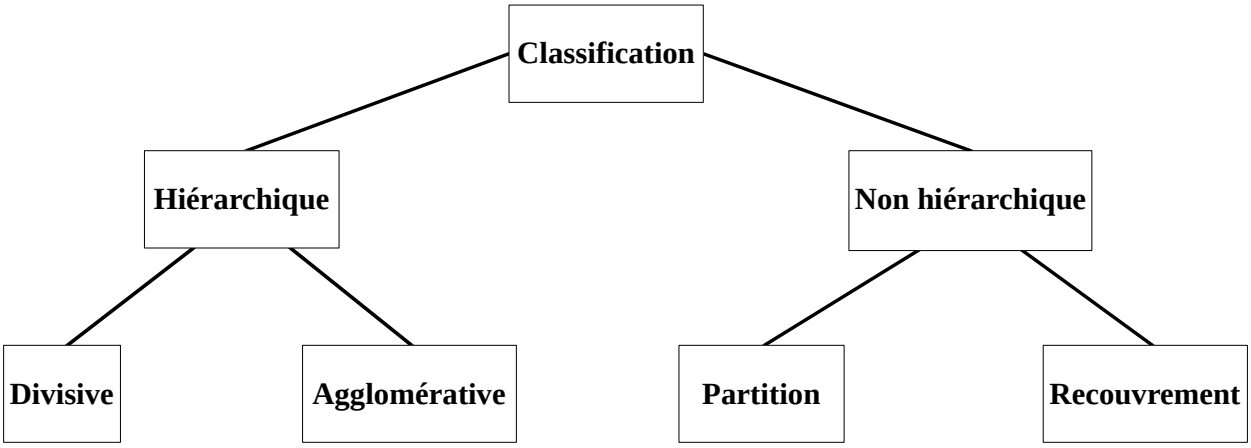
Dans une méthode *descendante* ou *divisive*, on part de l'ensemble de tous les éléments que l'on fractionne en un certain nombre de sous-ensembles. Ces derniers sont eux-même fractionnés récursivement jusqu'à ce que l'on arrive aux éléments individuels (Règne → Embranchement → Classe → … → Genre → Espèce)

Au contraire, dans un méthode *ascendante* ou *agglomérative*, on part des éléments individuels que l'on regroupe en sous-ensembles avant d'appliquer récursivement les regroupements sur les sous-ensembles jusqu'à obtenir un seul ensemble contenant tous les éléments :
(Espèce → Genre → … → Embranchement → Règne)

Méthodes recouvrantes et partition

En classification *non hiérarchique,* on peut considérer que chaque élément ne fait partie que d'*un* sous-ensemble, on parle alors de *partition*. On peut aussi considérer que chaque élément en fait partie de *plusieurs*, en attribuant une probabilité d'appartenance à chaque groupe et on parle alors de classification *recouvrante*.

CLASSIFICATION DES MÉTHODES DE CLASSIFICATION



TYPES DE MESURES

Chaque élément est caractérisé par un *ensemble de mesures*. Ces mesures peuvent être de différents types :

Données numériques

Valeur absolue (longueur, température)

Valeur relative (%)

Données nominales

Couleur des yeux (noir, vert, brun, …).

Pas de comparaisons possibles autres que $A = B$ ou $A \neq B$

Données ordinales

Grade à l'armée ; Petit, moyen, grand. *On peut définir un ordre $A < B < C$ …*

Données binaires

Cas particulier de variables nominales ou ordinales lorsqu'on a deux valeurs possibles

Sexe, présence ou absence d'une caractéristique

Données conditionnelles, liées à la présence d'autres mesures

Données manquantes

REPRÉSENTATION DES DONNÉES

Les données *brutes* peuvent être représentées par une matrice $\boldsymbol{X}$ à n lignes et p colonnes. Chaque ligne correspond à l'un des n éléments qu'on désire classer, caractérisé par p mesures.

Matrice de données brutes :

		Mesures					
		1	2	…	j	…	p
Éléments	1	x_{11}	x_{12}		x_{1j}		x_{1p}
	2	x_{21}	x_{22}		x_{2j}		x_{2p}
	…						
	i	x_{i1}	x_{i2}		x_{ij}		x_{ip}
	…						
	n	x_{n1}	x_{n2}		x_{nj}		x_{np}

Les colonnes ne sont pas forcément du même type et il peut y avoir des *entrées vides* correspondant aux entrées *manquantes*.

ÉTAPES DE LA CLASSIFICATION

1) *Sélection* d'un échantillon de n éléments parmi lesquels on cherche des groupes

2) *Mesure* des éléments.Effectuer p mesures sur chaque élément pour obtenir les *données brutes*.

3) *Construction* d'une matrice de *dissimilarités* $\boldsymbol{D}$, de taille $n \times n$. L'élément d_{rs} mesure la dissimilarité qui existe entre les éléments r et s. Tous les d_{rs} sont du *même type*, généralement un nombre réel avec $d_{rr} = 0$ et $d_{rs} \geq 0$.

Parfois les dissimilarités peuvent être booléennes : $d_{rs} = \begin{cases} 0 & \text{Les éléments } r \text{ et } s \text{ sont similaires} \\ 1 & \text{Les éléments } r \text{ et } s \text{ sont différents} \end{cases}$

4) *Choix d'un type de classification* (hiérarchique, partition, …) avec éventuellement des contraintes additionnelles : nombre de groupes à former, nombre maximum d'éléments dans un groupe, …

5) *Choix d'un critère* pour mesurer l'*homogénéité* ou l'*hétérogénéité* des groupes : attribuer une *mesure de qualité* pour chaque classification possible.

6) *Choix* ou conception d'un *algorithme* pour la résolution du problème ainsi défini.

CALCUL DES DISSIMILARITÉS

Selon le type des données, la manière de construire la matrice des dissimilarités varie.

Données binaires

Exemple : 0 1 0 0 1 0 0 1 1 objet *r*

1 1 0 1 0 1 1 0 1 objet *s*

Variable n^{o} 1 2 3 4 5 6 7 8 9

Sommaire des mesures entre *r* et *s*.

		Objet *r*	
		0	1
Objet *s*	0	1 = *a*	2 = *b*
	1	4 = *c*	2 = *d*

Quelques possibilités de définition de la dissimilarité :

Distance de Hamming : $d_{rs} = \frac{b + c}{a + b + c + d}$

Coefficient de Jaccard : $d_{rs} = \frac{b + c}{a + b + c}$. Propriété : c'est une distance ($d_{rt} \leq d_{rs} + d_{st} \forall r, s, t$)

Mesures liées aux probabilités : $1 - d_{rs} = \frac{ad - bc}{(a + b)(a + b + c + d)}$, $1 - d_{rs} = \frac{ad - bc}{(a + b)(a + c)}$

Approches géométriques : $1 - d_{rs} = \frac{a}{\sqrt{(a + b)(a + c)}}$ covariance

$1 - d_{rs} = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}$ coefficient de corrélation

CALCUL DE DISSIMILARITÉS (2)

Données nominales

Remplacer les *e* états possibles d'une variable d'un type énumératif non ordonné par *e* variables booléennes et se ramener au cas des données binaires.

Données ordinales

Définir une matrice de *coefficients de discordance* (δ_{pq}) entre les états *p* et *q* avec :

$\delta_{eq} > \delta_{e-1\,q} > \delta_{e-2\,q} > \ldots > \delta_{qq} = 0$ et $\delta_{p1} > \delta_{p2} > \delta_{p3} > \ldots > \delta_{pp} = 0$

		État de l'objet *r* (*j*$^{\text{ième}}$ mesure)					
		1	2	…	*q*	…	*e*
État de l'objet *s* (*j*$^{\text{ième}}$ mesure)	1	0					
	2		0				
	…						
	p	δ_{p1}			δ_{pq}		
	…						
	e	$\delta_{e1}=1$			δ_{eq}		0

La dissimilarité d_{rs} entre les objets *r* et *s* sera la somme (pondérée) des coefficients de discordances pour chaque mesure.

CALCUL DE DISSIMILARITÉS (3)

Données numériques

Distance de Manhattan : $d_{rs} = \sum_{j=1}^{p} |x_{rj} - x_{sj}|$ *Distance euclidienne* : $d_{rs} = \sqrt{\sum_{j=1}^{p} (x_{rj} - x_{sj})^2}$

Distance de Minkowski : $d_{rs} = \sqrt[\lambda]{\sum_{j=1}^{p} |x_{rj} - x_{sj}|^{\lambda}}$

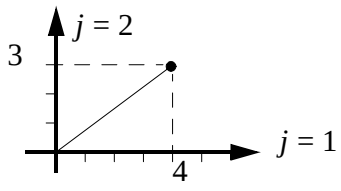
$\lambda = 1$: distance de Manhattan,

$\lambda = 2$: distance euclidienne, $\lambda = \infty$: maximum de la différence entre deux coordonnées.

$\lambda = 1 : d_{rs} = 4 + 3 = 7$

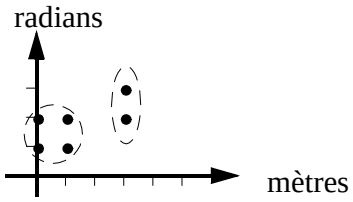
$\lambda = 2 : d_{rs} = \sqrt{16 + 9} = 5$

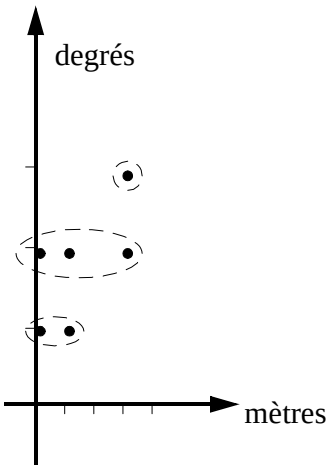
$\lambda = \infty : d_{rs} = \max(4, 3) = 4$



STANDARDISATION DES DONNÉES

Dans les calculs de dissimilarités présentés plus haut, on a fait l'hypothèse que toutes les mesures avaient la *même importance* et étaient basées sur la *même unité*, ce qui n'est pas forcément vrai. La structure des groupes peut être fortement influencée par les unités dans lesquelles les mesures ont été effectuées :





STANDARDISATION DES DONNÉES (2)

Pour éviter qu'une mesure prenne beaucoup plus d'importance que les autres, il convient de *normaliser* les unités.

Dans le cas d'une *distance de Manhattan*, il convient de diviser chaque composante par la différence *maximale* entre deux valeurs de cette composante. Dans le cas d'une *distance euclidienne*, il convient de diviser par l'*écart-type* des valeurs de chaque composante. De plus, il se peut que l'on désire donner explicitement un *poids* plus grand à une mesure qu'à une autre. On associera donc un poids w_j à chaque mesure, poids qui prendra en considération l'importance subjective que l'on veut donner à la $j^{\text{ième}}$ mesure divisée par un facteur de normalisation.

Exemple : on veut donner 4 fois plus d'importance à la 1ère mesure qu'à la 2e et utiliser des distances euclidiennes.

On aura : $w_1 = \dfrac{4}{\frac{1}{n} \cdot \left(\sum (x_{i1})^2 - \dfrac{\left(\sum x_{i1}\right)^2}{n} \right)}$, $w_2 = \dfrac{1}{\frac{1}{n} \cdot \left(\sum (x_{i2})^2 - \dfrac{\left(\sum x_{i2}\right)^2}{n} \right)}$, $d_{rs} = \sqrt{\sum_{j=1}^{p} w_j (x_{rj} - x_{sj})^2}$

MESURES DE DIFFÉRENTS TYPES

Très souvent, les données brutes sont de *plusieurs types*. Par exemple, un sol peut être caractérisé par son acidité (Ph, numérique), sa couleur (ordinal), le type de pierres présentes (nominal) et la présence de vers (binaires).

Si la *majorité des mesures sont d'un type*, on peut essayer de ramener toutes les données à ce type :

Numérique	→	Booléen	acide, basique
Booléen	→	Numérique	0, 1
Ordinal	→	Numérique	rang
Numérique	→	Nominal	classes non ordonnées
Numérique	→	Ordinal	classes ordonnées

S'il n'y a *pas de type prépondérant*, il faut passer par la définition d'un *poids à attribuer à chaque critère*.

VALEURS MANQUANTES

Si une mesure *k* est *manquante* pour l'élément *r* (et/ou *s*) on peut l'*ignorer* dans le calcul des dissimilarités :

$$d_{rs} = \sqrt{\sum_{j=1, j \neq k}^{p} w_j (x_{rj} - x_{sj})^2}$$

Une autre technique consiste à *remplacer* la valeur manquante par la *moyenne des autres valeurs* ou bien par la valeur de l'élément le plus similaire.

CRITÈRES DE SÉPARATION

On suppose dès maintenant que l'on dispose d'une matrice carrée dont les entrées mesurent la dissimilarité qu'il y a entre chaque paire d'éléments. La *séparation* d'un groupe *G* par rapport aux autres groupes peut être mesuré de diverses manières :

$$\text{séparation}(G) = min_{r \in G, s \notin G}(d_{rs})$$

$$\text{coupe}(G) = \sum_{r \in G} \sum_{s \notin G} d_{rs} ; \qquad \text{coupe normalisée}(G) = \frac{\sum_{r \in G} \sum_{s \notin G} d_{rs}}{|G|(n - |G|)}$$

Un *critère global* de mesure de *qualité* d'une classification, basé sur les séparations des groupes, peut être défini de diverses manière, selon que l'on se focalise sur le groupe le moins bien séparé ou sur la séparation moyenne de groupes :

$$max_G(\text{séparation}(G)), \qquad max_G(\text{coupe}(G))$$

$$max(\sum_G \text{séparation}(G)), \qquad max(\sum_G \text{coupe}(G))$$

CRITÈRES D'HOMOGÉNÉITÉ

L'homogénéité des éléments à l'intérieur d'*un* groupe *G* peut également être définie de diverses manières :

diamètre(G) = $max_{r,\, s \in G}(d_{rs})$

rayon(G) = $min_{r \in G}(max_{s \in G}(d_{rs}))$

étoile(G) = $min_{r \in G}(\sum_{s \in G} d_{rs})$ étoile normalisée(G) = $\frac{min_{r \in G}(\sum_{s \in G} d_{rs})}{|G| - 1}$

clique(G) = $\sum_{r \in G} \sum_{s \in G} d_{rs}$ clique normalisée(G) = $\frac{\sum_{r \in G} \sum_{s \in G} d_{rs}}{|G|(|G| - 1)}$

Le critère *global* de qualité d'une classification, basé sur l'homogénéité consistera à *minimiser* le *plus grand diamètre* (rayon, étoile, …) ou la *somme* des diamètres (rayon, …) pour l'*ensemble des groupes.*

CRITÈRES BASÉS SUR DES CENTROÏDES

Dans le cas de mesures dans un *espace euclidien*, on peut mesurer l'homogénéité de *G* par rapport à un centre $\bar{c}$ ne faisant pas partie des éléments. $\bar{c}$ peut être défini comme le *centre de gravité* des éléments de *G* :

$$\bar{c} = \frac{1}{|G|} \sum_{i \in G} (x_{i1}, x_{i1}, \ldots, x_{ip}) = \frac{1}{|G|} \sum_{i \in G} \vec{x}_i$$

Le critère d'homogénéité basé sur la *somme des carrés* des distances au centre $\bar{c}$ est très répandu :

$$\mathrm{SC}(G) = \sum_{i \in G} \left\| \vec{x}_i - \bar{c} \right\|^2$$

Pour ce critère, le centre de gravité $\bar{c}$ est le point minimisant la valeur de SC(G). Pour le critère basé sur la somme des distances :

$$\mathrm{SD}(G) = \sum_{i \in G} \left\| \vec{x}_i - \tilde{c} \right\|$$

le centre $\tilde{c}$ minimisant la valeur de la somme des distances ne peut plus s'exprimer analytiquement comme pour $\bar{c}$; il est nécessaire d'*estimer numériquement* la position optimale de $\tilde{c}$ pour un groupe *G.*

CLASSIFICATION HIÉRARCHIQUE

Méthode agglomérative

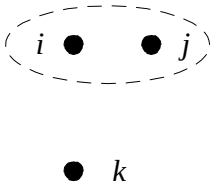
Idée générale : créer tout d'abord *n* groupes ne contenant qu'*un* élément. À chaque étape, *fusionner* les *deux* groupes les *plus proches* et recalculer une dissimilarité entre ce nouveau groupe et les autres.

Algorithme général

```
I = {1 .. n}          -- indices des groupes
G_i = {i}, n_i = 1    -- groupes et nombre d'éléments par groupe
Calculer d_ij pour tout i, j ∈ I
Pour k = 1 .. n - 1 répéter :
   Trouver un d_ij minimum (i, j ∈ I)

   -- Fusionner les groupes i et j.
   Poser G_i ← G_i ∪ G_j; n_i ← n_i + n_j

   -- Recalculer les distances
   d_ki ← α_i d_ik + α_j d_jk + β d_ij + δ|d_ik - d_jk| pour tout k ∈ I\{i,j}
   d_ik ← d_ki pour tout k ∈ I\{i,j}
   -- Mettre à jour l'ensemble des indices
   I ← I\{j}
```



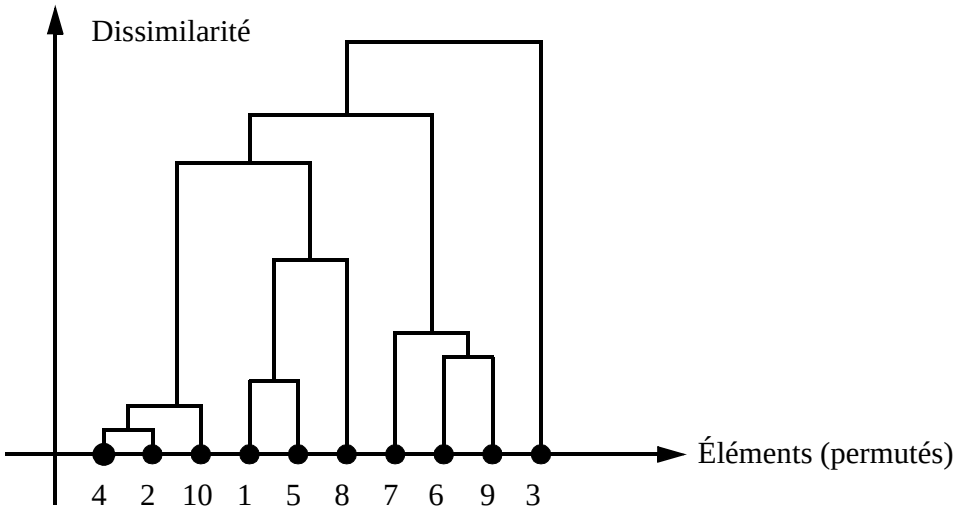
STRATÉGIES DE RECALCUL DES DISTANCES

Dans l'algorithme général précédent, plusieurs *stratégies* ont été proposées pour le calcul des distances (ligne : $d_{ki} \leftarrow \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \delta|d_{ik} - d_{jk}|$) une fois que *deux groupes ont été fusionnés*. Ces stratégies peuvent être caractérisées par diverses manières de calculer α_i, α_j, β, et δ :

	α_i	α_j	β	δ
Lien simple (plus proche voisin)	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Lien complet (voisin le plus éloigné)	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Lien moyen	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0
Centre	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$-\frac{n_i n}{n_i + n_j}$	0
Somme des carrés incrémentale (Ward)	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{n_j + n_k}{n_i + n_j + n_k}$	$-\frac{n_k}{n_i + n_j + n_k}$	0

DENDROGRAMMES

Une fois qu'une classification hiérarchique a été trouvée, il est possible de la représenter *graphiquement* au moyen de *dendrogrammes* (diagramme en arbre) :



Durant la construction, il faut *réordonner les éléments* de sorte qu'il n'y ait pas de croisement entre ligne horizontale et verticale.

CLASSIFICATION GLOBALE

Problème de la *p*-médiane

Le problème de la *p*-médiane est très souvent utilisé pour procéder à une classification globale. Il peut être présenté ainsi : On cherche *p* centres parmi les *n* éléments, tels que si l'on rattache chaque élément à son centre le plus proche, la somme totale des dissimilarités est minimale. Le critère global d'optimisation est donc la minimisation de la somme des étoiles.

Exemple graphique :

Deux classifications possibles avec $p = 2$, selon les centres choisis.

FORMULATION MATHÉMATIQUE :

1) Étant donné une matrice $\boldsymbol{D}$ $(n \times n)$ de dissimilarités, trouver p indices $J = \{j_1, j_2, \ldots j_p\}$

minimisant : $\sum_{i=1}^{n} min_{j \in J}(d_{ij})$

2) Introduire des variables indicatrices $y_j = \begin{cases} 1 & j \text{ est sélectionné} \\ 0 & j \text{ non sélectionné} \end{cases}$ et $x_{ij} = \begin{cases} 1 & i \text{ est rattaché à } j \\ 0 & i \text{ non rattaché à } j \end{cases}$.

minimiser	$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}x_{ij}$		Somme des étoiles
sous	$\sum_{j=1}^{n} x_{ij} = 1$	$\forall i$	On doit rattacher chaque élément à un centre
contraintes	$x_{ij} \leq y_j$	$\forall i, j$	On ne peut pas allouer un élément à un centre non sélectionné
	$\sum_{j=1}^{n} y_j = p$		On doit ouvrir p centres
	$x_{ij} \in \{0,1\}$	$\forall i, j$	
	$y_j \in \{0,1\}$	$\forall i$	

COMPLEXITÉ

Le problème de la p-médiane est *difficile* : on ne connaît pas d'algorithme polynomial pour le résoudre (c'est-à-dire en $O(n^c)$ avec une constante c ne dépendant ni de n, ni de p). Par contre, si p est fixé, il existe un algorithme polynomial évident, mais qui ne peut être mis en œuvre que pour des valeurs de p très petites :

```
Pour j1 = 1 .. n - p + 1
   Pour j2 = j1 .. n - p + 2
      ...
         Pour jp = jp-1 .. n
            Ouvrir les centres j1, j2, ... jp

            Trouver le rattachement optimum des éléments

            Calculer la valeur de l'objectif global

            Si la solution la meilleure trouvée jusqu'ici, la mémoriser
```

ALGORITHMES HEURISTIQUES DE RÉSOLUTION

1) *K*-Means (dans ce nom, le *K* fait référence au nombre de centre i.e. ≡ *p*)

Demander à un expert de placer les *p* centres. En l'absence d'expert, utiliser une procédure constructive simple (par exemple, placer les centres aléatoirement)

Répéter
   Allouer les éléments au centre le plus proche
   Évaluer la qualité de la solution
   Pour chaque groupe *G* d'éléments alloués au même centre, trouver le meilleur centre possible parmi les éléments de *G*
Tant que l'on réussi à améliorer la qualité de la solution.

2) *K*-Means itéré

Répéter *I* **fois**
   Placer les centres aléatoirement
   Appliquer *K*-Means
   Mémoriser la solution si elle est meilleure
Tant que l'on réussi à améliorer la qualité de la solution.
Retourner la meilleure solution trouvée.

3) *K*-Means avec repositionnement

Placer les *p* centres aléatoirement (= *solution_courante*);
Poser *meilleure_solution* := *solution_courante*;

Répéter
   Pour *i* = 1..*p*, **répéter**
      Pour *j* = 1..*n*, **s'**il n'y a pas de centre en *j*, **répéter**
         solution_courante := *meilleure_solution*
         Déplacer le centre *i* en *j* dans *solution_courante*
         Répéter -- appliquer *K*-Means à *solution_courante*
            Allouer les éléments au centre le plus proche
            Évaluer la qualité de *solution_courante*
            Pour chaque groupe *G* d'éléments alloués au même centre :
               Placer le centre optimalement dans *solution_courante*
         Tant que l'on a réussi à améliorer la qualité *solution_courante*

         Si *solution_courante* est meilleure que *meilleure_solution*
            meilleure_solution := *solution_courante*
         Fin_si
      Fin_pour *j*
   Fin_pour *i*
Tant que l'on a réussi à améliorer la qualité de *meilleure_solution*
Retourner *meilleure_solution.*