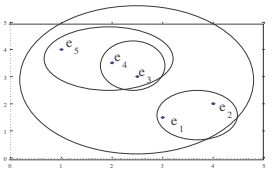
Méthodes en classification automatique

Classification hiérarchique

Yves Lechevallier
INRIA-Rocquencourt
78153 Le Chesnay Cedex
E_mail : Yves.Lechevallier@inria.fr

Structure classificatoire

Hiérarchie



1) $E \in H$

2) $\forall e \in E$ alors $\{e\} \in H$

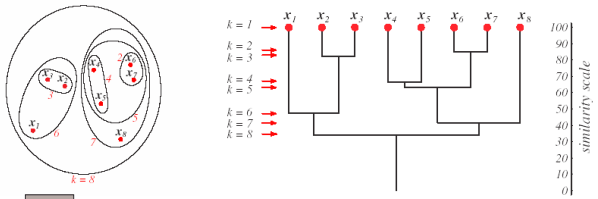
3) $\forall h, h' \in H$ on a :

$h \cap h' \neq \varnothing \Rightarrow h \subset h'$ ou $h' \subset h$

Classification hiérarchique

Diagramme de Venn sur des données bidimensionnelles

Dendrogramme

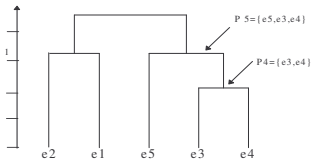


Méthodes hiérarchiques

En partant du tableau de données on calcule une distance entre les individus de E . Cette distance d vérifie :

1. $d(\mathbf{x},\mathbf{y}) = 0 \Rightarrow \mathbf{x} = \mathbf{y}$
2. $\forall \mathbf{x},\mathbf{y} \in D\ d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$ (symétrie)
3. $\forall \mathbf{x},\mathbf{y},\mathbf{z} \in D\ d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$ (inégalité triangulaire)

Les méthodes hiérarchiques ont pour objectif de construire une suite de partitions emboîtées appelée *hiérarchie*. La représentation graphique de ces hiérarchies se fait par un *arbre hiérarchique* ou *dendrogramme*.



Algorithmes hiérarchiques

Construction d'un dendrogramme

Complexité

➢à partir du bas c'est-à-dire à partir des feuilles terminales de l'arbre. Dans ce cas, on agrège, deux par deux, les classes les plus proches. De proche en proche ce processus est utilisé jusqu'à l'obtention d'une seule classe.

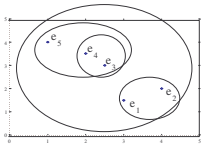
N^2

➢à partir du haut c'est-à-dire en procédant par division successive de l'ensemble jusqu'à obtenir un seul individu par classe. On obtient ainsi les feuilles terminales de l'arbre .

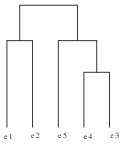
$2^{N-1}-1$

Hiérarchie indicée

Hiérarchie H



Hiérarchie indicée (H,f)



1) $E \in H$

2) $\forall e \in E$ alors $\{e\} \in H$

3) $\forall h, h' \in H$ on a :

$h \cap h' \neq \varnothing \Rightarrow h \subset h'$ ou $h' \subset h$

$f : H \to \Re^+$

(1) $f(h) = 0$ si et seulement si $card(h)=1$

(2) $\forall h, h' \in H \; h \subset h'$ et $h \neq h' \Rightarrow f(h) < f(h')$

Équivalence entre hiérarchies indicées et ultramétriques

1) δ est une distance ultramétrique

$$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \; \delta(\mathbf{x}, \mathbf{y}) \leq \mathrm{Max}\{\delta(\mathbf{x}, \mathbf{z}), \delta(\mathbf{z}, \mathbf{y})\}$$

2) Construction d'une ultramétrique induite par une hiérarchie indicée

(H,f) une hiérarchie indicée

$$\delta(\mathbf{x}, \mathbf{y}) = Min\{f(h) / h \in H, \mathbf{x} \text{ et } \mathbf{y} \in h\}$$

δ est une ultramétrique

$$\Phi : (H, f) \to \delta$$

3) Construction d'une hiérarchie indicée à partir d'une ultramétrique

$$\forall \mathbf{x}, \mathbf{y} \quad \mathbf{x} R_\alpha \mathbf{y} \Leftrightarrow \delta(\mathbf{x}, \mathbf{y}) \leq \alpha \;\; \alpha > 0$$

✓R_α est une relation d'équivalence

✓Les classes d'équivalences forment une hiérarchie

✓f(h) est le diamètre de la classe h

$$\Psi : \delta \to (H, f)$$

$$\Phi = \Psi^{-1}$$

Recherche d'une hiérarchie indicée en terme d'optimisation

U ensemble de ultramétriques

Recherche δ de U optimisant $\Delta(d,\delta)$

$$\Delta(d, \delta_\alpha) = \min_{\delta \in U} \left[\sum_{\mathbf{x},\mathbf{y}} |d(\mathbf{x}, \mathbf{y}) - \delta(\mathbf{x}, \mathbf{y})|^\alpha \right]^{1/\alpha}$$

Si $\alpha = 2$ c'est une optimisation au sens des moindres carrés

Ultramétriques sous-dominante et sur-dominante

sous-dominante $\delta_s(\mathbf{x},\mathbf{y}) = Sup\left[\delta(\mathbf{x},\mathbf{y}) / \delta \in U, \delta \le d\right]$

δ_s est une ultramétrique et elle est unique

sur-dominante $\overline{\delta}(\mathbf{x},\mathbf{y}) = Min\left[\sum_{\mathbf{x},\mathbf{y}} \left| d(\mathbf{x},\mathbf{y}) - \delta(\mathbf{x},\mathbf{y}) \right| / \delta \in U, \delta \ge d\right]$

Elle n'est pas unique

Remarque :

$\ddot{\delta}(\mathbf{x},\mathbf{y}) = Inf\left[\delta(\mathbf{x},\mathbf{y}) / \delta \in U, \delta \le d\right]$ $\ddot{\delta}$ n'est pas forcement une ultramétrique

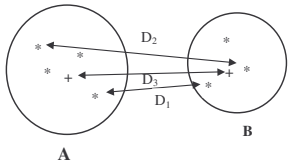
Les indices d'agrégation entre les classes

lien minimum

$$D_1(A,B) = \min_{\substack{a\in A \\ b\in B}} d(a,b)$$

lien maximum

$$D_2(A,B) = \max_{\substack{a\in A \\ b\in B}} d(a,b)$$



l'augmentation de l'inertie ou indice de WARD

$$D_3(A,B) = I(A\cup B) - I(A) - I(B) = \frac{\mu(A)\mu(B)}{\mu(A)+\mu(B)} d^2(g_A, g_B)$$

g_A est le centre de gravité de la classe A

μ correspond à la pondération des classes

Relation entre *f* et *D*

f est un indice sur la hiérarchie *H*, *D* est un indice d'agrégation entre classes

$$\forall h_1, h_2 \in H \quad f(h_1 \cup h_2) = D(h_1, h_2)$$

Pour les indices *D* courants (*H*,*f*) est une hiérarchie indicée (il n'y a pas d'inversion)

Sinon on peut utiliser

$$\forall h_1, h_2 \in H \quad f(h_1 \cup h_2) = Max\left\{D(h_1,h_2), f(h_1), f(h_2)\right\}$$

Dans ce cas (*H*,*f*) est toujours une hiérarchie indicée

Construction de l'ultramétrique

A partir de cette hiérarchie indicée (*H*,*f*) , on définit, une distance ultramétrique δ entre les individus de *E*.

$$\delta(\mathbf{z}_i, \mathbf{z}_m) = \min_{h\in H}\left\{f(h) / e_i \in h \text{ et } e_m \in h\right\}$$

De manière constructive on a:

$$\forall e_i \in h_1, e_j \in h_2 \text{ on a } \delta(\mathbf{z}_i, \mathbf{z}_j) = D(h_1,h_2) = f(h_1 \cup h_2)$$

Donc, plus le palier regroupant un ensemble d'individus se trouve dans le bas de l'arbre, plus la valeur de l'ultramétrique δ entre ces individus de cet ensemble est petite

Algorithme de la classification ascendante hiérarchique CAH

***(a)* initialisation**

On se donne au départ la partition constituée de N classes

$$Q^{(0)} = (Q_1^{(0)}, \cdots, Q_N^{(0)}) \text{ avec } Q_i^{(0)} = \{e_i\}$$

On se donne un indice d'agrégation $D : \mathrm{P}(E) \times \mathrm{P}(E) \to \Re^+$ qui vérifie $\forall e_i, e_m \in E \quad D(\{e_i\}, \{e_m\}) = d(\mathbf{z}_i, \mathbf{z}_m)$

***(b)* Étape agrégative**

Construire une nouvelle partition $Q^{(N-K)}$ contenant K classes à partir de la partition $Q^{(N-K-1)}$ contenant $K+1$ classes en réunissant les deux classes de $Q^{(N-K-1)}$ les plus proches au sens de la mesure d'agrégation D.

(c) Recommencer l'étape **(b)** jusqu'à obtenir une seule classe, c'est-à-dire la partition grossière.

La formule de récurrence de Lance et Williams

Il est nécessaire de recalculer l'indice d'agrégation entre la nouvelle classe ainsi formée et les autres classes de la partition. Lance et Williams en 1967 ont proposé, lors du regroupement des deux classes , la formule de récurrence suivante :

$$A, B, C \in \mathrm{P}(E) D(C, A \cup B) = \alpha_1 D(C, A) + \alpha_2 D(C, B) + \alpha_3 D(A, B) + \alpha_4 \left| D(C, A) - D(C, B) \right|$$

lien minimum $\quad \alpha_1 = \alpha_2 = 1/2, \alpha_3 = 0, \alpha_4 = -1/2$

lien maximum $\quad \alpha_1 = \alpha_2 = 1/2, \alpha_3 = 0, \alpha_4 = 1/2$

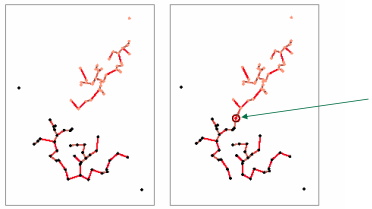
l'augmentation de l'inertie ou indice de WARD

$$\alpha_1 = \frac{\mu(C)+\mu(B)}{\mu(F)}, \alpha_2 = \frac{\mu(C)+\mu(A)}{\mu(F)}, \alpha_3 = -\frac{\mu(C)}{\mu(F)}, \alpha_4 = 0$$

avec $F = A \cup B \cup C$

Lien minimum

Deux distributions gaussiennes. L'algorithme hiérarchique du lien minimum reconnaît bien les deux classes



L'ajout d'un point marqué d'un cercle rouge perturbe beaucoup le résultat

Les méthodes descendantes polythétiques

Ces méthodes sont généralement définies à partir d'un tableau de dissimilarités. Une des plus anciennes méthodes est celle de MacNaughton-Smith (1964).

Ces méthodes cherchent généralement à optimiser un critère local, c'est-à-dire le critère de la bipartition et non celui de la partition induite, sans faire une énumération complète de toutes les bipartitions.

Les méthodes descendantes monothétiques

Ces méthodes imposent aux individus d'une même classe de posséder au moins une caractéristique en commun. Cette propriété est généralement obtenue en divisant une classe en fonction d'une variable et d'une dichotomie des valeurs de cette variable.

L'intérêt de cette approche est la facilité d'interprétation des classes obtenues. En effet, chaque classe est définie par une conjonction de caractéristiques traduisant une condition nécessaire et suffisante d'appartenance à la classe.

Les méthodes descendantes monothétiques

La première méthode divisive de type monothétique a été proposée par Williams et Lambert (1959, 1960, 1961)sur des tableaux binaires. Cette méthode a été étendue au cas des variables qualitatives (Volle, 1976), créant un nouveau problème de complexité, le nombre de dichotomies augmentant de manière exponentielle avec le nombre de modalités.

Depuis, les méthodes monothétiques de classification divisive ont été étudiées en intelligence artificielle dans le cadre des méthodes de classification conceptuelle

DIV : une méthode divisive

Cette méthode (Chavent 1997, 1998 divise à chaque étape une classe en fonction d'une question binaire et du critère d'inertie.

Dans le cas de variables quantitatives, la méthode utilise soit la distance euclidienne usuelle.

Dans le cas de variables qualitatives, la méthode utilise soit la distance euclidienne usuelle soit la distance du khi-deux sur le tableau disjonctif complet.

A chaque étape, la méthode définit la question binaire qui induit la bipartition d'inertie intra-classe minimum.

Questions binaires

variable continue

$[X > 3.5]$?

Variable qualitative

$[X \in \{m_1,\ldots,m_h\}]$?

•Dans le cas d'une variable continue on évalue toutes coupures possibles c'est-à-dire au maximum *n-1*

• Pour une variable qualitative ordonnée *Y*, on évalue ainsi au maximum *m*-1 bipartitions

•Dans le cas d'une variable qualitative non ordonnée, on se heurte vite à un problème de complexité, le nombre de dichotomies du domaine d'observation étant alors égal à 2^{m-1}-1.

Critère d'évaluation

Soit $P=(P_1,\ldots,P_K)$ une partition en K classes

Critère d'évaluation $W(P)$ doit être additif $W(P)=\sum_{P_k \in P} w(C_k)$

Exemple : Inertie intra-classe

La réduction du critère d'évaluation revient à maximiser le gain $\Delta(Q)$ associé à la question binaire Q de découper la classe C et deux classes C_1 et C_2

$$\Delta(Q^*)=\max_{Q \in B} |w(C)-w(C_1)-w(C_2)|$$

B étant l'ensemble des questions binaires admissibles

Algorithme divisif (récursif)

Étape 1: Tous les objets dans la même classe C

Étape 2: Diviser successivement chaque classe C en deux classes (C_1,C_2) en fonction du critère de l'inertie intra-classes.

étape 2.1: pour chaque variable X , trouver la coupure s qui maximise

$$\Delta(X,s/C)=|w(C) - w(C1)-w(C2)|$$

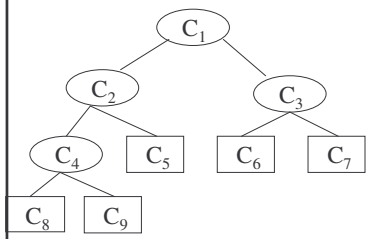
étape 2.2: choisir la variable X^* et la coupure s

$$\Delta(X^*,s^*/C)=max\ \Delta(X,s/C)$$

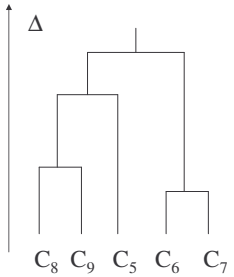
étape 3: diviser la classe C en (C_1,C_2)

Arbre de décision/ Hiérarchie indicée

Pas d'ordre de découpage



Ordre de construction



Output : résultats

- **Les *K* partitions de notre ensemble d'individus**
- **Hiérarchie indicée**
- **Chaque nœud représente une classe**
- **Chaque classe peut être décrite par une règle**

Les données

Ces deux méthodes ont été utilisées lors d'une analyse menée au CE.R.I.E.S. L'objectif de cette analyse était de proposer une typologie fiable de la peau humaine saine reposant sur un petit nombre de caractères cutanés pertinents.

Les données ont été recueillies entre avril et mai 1996, sur 212 femmes volontaires d'Ile-de-France présentant une peau saine et d'âge compris entre 20 et 50 ans.

Ces données résultent d'un examen clinique appréciant 17 caractéristiques de la peau de la joue, évaluées sur des échelles qualitatives.

Ces signes cliniques peuvent être visuels comme « aspect gras » ou encore « grain irrégulier de la peau ». Ils peuvent également être tactiles comme « toucher rêche » ou encore « Incapacité à rosir même au pincement ». Ces variables sont toutes binaires ou ordinales.

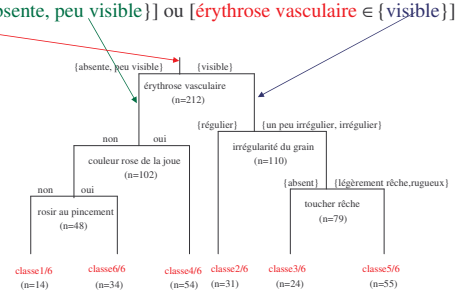
Méthodes de classification divisives et segmentation non supervisées: recherche d'une typologie de la peau humaine; Marie Chavent, Christiane Guinot, Yves Lechevallier, Michel Tenenhaus· RSA, 1999

Arbre hiérarchique

la première question binaire est :

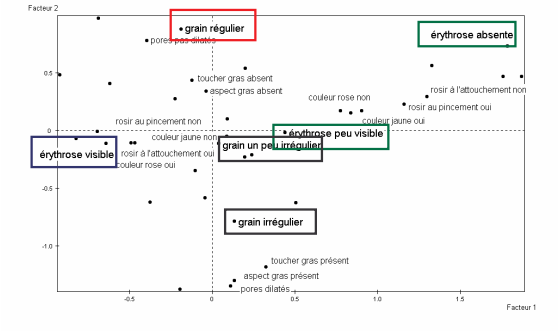
[érythrose vasculaire ∈ {absente, peu visible}] ou [érythrose vasculaire ∈ {visible}]

Les 212 femmes sont donc divisées en 102 femmes dont l'érythrose vasculaire est absente ou peu visible et 110 femmes dont l'érythrose vasculaire est visible.



Analyse factorielle

La première coupure est associée à l'axe 1 et la seconde à l'axe 2



Évolution des classes

Visualisation de la partition en trois classes et des centres de gravité des classes des partitions de 2 à 6 classes