MATHEMATIQUES APPLIQUEES de BORDEAUX

Critères de Rand asymétriques
Application en chimie organique

M. Chavent - C. Lacomblez - B. Patouille

MATHEMATIQUES APPLIQUEES de BORDEAUX

Un problème de validation externe

Ω Population étudiée

P : partition *a priori* *(experte)*

Ensemble {Q} de partitions *(classification automatique)*

Trouver la partition *Q* la plus en « accord » avec *P*

MATHEMATIQUES APPLIQUEES de BORDEAUX

Comparer deux partitions *P* et *Q* d'un même ensemble

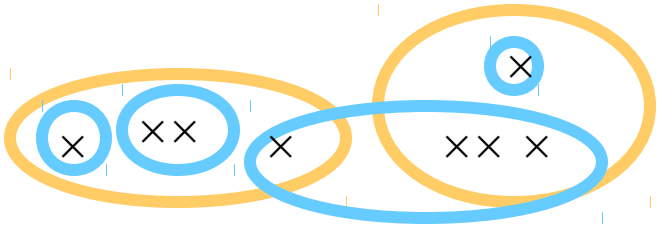
Critères symétriques :

- Critère de Rand (1971)
- Critère de Rand corrigé (1985)

P et *Q* ont le même nombre de classes

Comparer deux partitions *P* et *Q* d'un même ensemble

P et Q n'ont pas le même nombre de classes :



Dans quelle mesure P est-elle plus « fine » que Q ?

Critères asymétriques

MATHEMATIQUES APPLIQUEES de BORDEAUX

Plan de l'exposé

- Critère de Rand asymétrique
- Critère de Rand asymétrique corrigé
- Application en botanique

Notations

	Q_1	…	Q_j	…	Q_l
P_1			⋮		
⋮					
P_i		…	n_{ij}	…	
⋮					
P_k			⋮		

Tableau de contingence

Partition P \ Partition Q	même classe	≠ classes
même classe	a	b
≠ classes	c	d

Tableau accords-désaccords

Formules de passage

MATHEMATIQUES APPLIQUEES de BORDEAUX

Critère de Rand

$$R(P,Q) = \begin{cases} \dfrac{a+d}{a+b+c+d} \\ 1 + \dfrac{2\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_{i}\binom{n_{i.}}{2} + \sum_{j}\binom{n_{.j}}{2}\right]}{\binom{n}{2}} \end{cases}$$

- $R(P,Q) = R(Q,P)$ = *pourcentage du nombre d'accords* ($\mathbf{a+d}$)
- $R \in [0,1]$ *et* $R=1$ *si* $\forall i \, \exists j$ *tq* $P_i = Q_j$
- $R = P(A \cap B) = |A \cap B| / |E|$ *où*
  - *E=ensemble des paires {x,y}*
  - *A={x et y classés ensemble selon P}*
  - *B={x et y classés ensemble selon Q}*
- *R évalue la règle* $A \Leftrightarrow B$

MATHEMATIQUES APPLIQUEES de BORDEAUX

Critère de Rand asymétrique

$$RA(P,Q) = \begin{cases} \dfrac{a+d+c}{a+b+c+d} \\ 1 + \dfrac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_{i}\binom{n_{i.}}{2}}{\binom{n}{2}} \end{cases}$$

- $RA(P,Q) \neq RA(Q,P)$ = *pourcentage du nombre d'accords* (**a+d** +**c**)
- $RA \in [0,1]$ *et* $RA=1$ *si* $\forall i \; \exists j$ *tq* $P_i \subseteq Q_j$
- $RA = 1 - P\left(A \cap \overline{B}\right) = 1 - \dfrac{2}{n(n-1)} \sum_{i=1}^{k} \sum_{1 \le s < t \le l} \left|P_i \cap Q_s\right| \left|P_i \cap Q_t\right|$

(Bel Mufty, 98)

- *RA évalue la règle* $A \Rightarrow B$

Problème d'échelle des critères *R* et *RA*

Quand a-t-on $R=0$ ou $RA=0$?

Critère de Rand corrigé (Hubert & Arabie, 85) d'espérance nulle lorsque les accords sont dus au hasard

$$Rc=\frac{R-E(R)}{1-E(R)}$$

Critère de Rand Asymétrique corrigé

$$RAc=\frac{RA-E(RA)}{1-E(RA)}$$

Propriété de l'espérance

Sous l'hypothèse nulle où P et Q sont tirées au hasard dans l'ensemble des partitions en k et l classes, les nombres n_j, n_i d'individus par classe étant fixés, alors N_{ij} (nombre d'individus dans $P_i \cap Q_j$) suit $H(n,n_j,n_i)$

nombre de paires dans $P_i \cap Q_j$

nombre de paires dans P_i

nombre de paires dans Q_j

$$E\left(\sum_{i,j}\binom{n_{ij}}{2}\right)=\frac{\sum_i\binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}}{\binom{n}{2}}$$

nombre de paires dans Ω

(Folkes & Mallow, 83)

MATHEMATIQUES APPLIQUEES de BORDEAUX

$$RAc = \frac{RA - E(RA)}{1 - E(RA)}$$

$$E(RA) = 1 + \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}^2} - \frac{\sum_i \binom{n_{i.}}{2}}{\binom{n}{2}}$$

Critère de Rand Asymétrique corrigé

$$RAc = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}}}{\sum_i \binom{n_{i.}}{2} - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}}}$$

Application en botanique.

Étude d'une famille de conifères (les Pinacées)

- 38 feuilles

- une variable « genre » nominale à 6 modalités :
les pins (*Pinus*), les sapins (*Abies*), les épicéas (*Picea*)
…

- 24 variables continues : les quantités (en pourcentage) de chacun des 24 acides gras présents dans la composition des feuilles

Tableau 1 : Extrait du tableau de données

	Genre	12:0	14:0	…	22:1
1	***Pinus***	**2,5**	**2,5**		**0**
…	…				
	Picea	**0,3**	**0,4**		**0**
…	…				
	Larix	**0**	**5,5**		**0**
…	…				
38	***Abies***	**5,5**	**2,0**		**0**

Une partition experte *P* en 6 classes des 38 feuilles

Plusieurs partitions *Q* en *j* <6 classes des 38 feuilles obtenues par classification automatique

MATHEMATIQUES APPLIQUEES de BORDEAUX

Plusieurs partitions *Q* obtenues par classification automatique

Classification hiérarchique de Ward sur coordonnées factorielle obtenues par:

- ACP sur les 24 variables continues transformée ou non par une transformation de Box-Cox :

  $y=(x^{\lambda}-1)/\lambda$ *si* $\lambda \neq 0$ *et* $y=ln(x)$ *sinon*

  ici $\lambda=½$ *et* $y=2(x^{½}-1)$

- ACP normée ou non normée.

MATHEMATIQUES APPLIQUEES de BORDEAUX

Les différentes options de classification envisagées pour des partitions en j=4,5,6 classes

	ACP normée	**ACP non normée**
Variables non transformées	Cas1/j	Cas2/j
Variables transformées	Cas3/j	Cas4/j

➡ 4×3 partitions générées avec SPAD

MATHEMATIQUES APPLIQUEES de BORDEAUX

Valeur des 4 critères pour les 12 partitions

Cas1: nonTr/N →

Cas2 : nonTr/nonN →

Cas3 : Tr/N →

Cas 4 : Tr/nonN →

	R	Rc	RA	RAc
Cas 1/ 3	**0.686**	**0.300**	**0.964**	**0.655**
Cas 1/ 4	**0.690**	**0.279**	**0.952**	**0.553**
Cas 1/ 5	**0.794**	**0.414**	**0.943**	**0.561**
Cas 2/ 3	**0.720**	**0.321**	**0.953**	**0.585**
Cas 2/ 4	**0.799**	**0.445**	**0.953**	**0.631**
Cas 2/ 5	**0.811**	**0.442**	**0.942**	**0.562**
Cas 3/ 3	**0.718**	**0.366**	**0.977**	**0.782**
Cas 3/ 4	**0.775**	**0.382**	**0.943**	**0.550**
Cas 3/ 5	**0.794**	**0.392**	**0.933**	**0.498**
Cas 4/ 3	**0.650**	**0.256**	**0.963**	**0.619**
Cas 4/ 4	**0.832**	**0.524**	**0.963**	**0.714**
Cas 4/ 4	**0.811**	**0.442**	**0.942**	**0.562**

MATHEMATIQUES APPLIQUEES de BORDEAUX

Analyse des résultats du tableau

Les critères symétriques *R* et *Rc* ont tendance à augmenter avec le nombre de classes indépendamment de la méthode utilisée

- Permettent éventuellement de choisir parmi deux partitions ayant le même nombre de classes celle qui « ressemble » le plus à la partition experte

- Ne semblent pas appropriés pour choisir parmi les 12 partitions

	R	Rc
Cas 1/ 3	**0.686**	**0.300**
Cas 1/ 4	**0.690**	**0.279**
Cas 1/ 5	**0.794**	**0.414**
Cas 2/ 3	**0.720**	**0.321**
Cas 2/ 4	**0.799**	**0.445**
Cas 2/ 5	**0.811**	**0.442**
Cas 3/ 3	**0.718**	**0.366**
Cas 3/ 4	**0.775**	**0.382**
Cas 3/ 5	**0.794**	**0.392**
Cas 4/ 3	**0.650**	**0.256**
Cas 4/ 4	**0.832**	**0.524**
Cas 4/ 4	**0.811**	**0.442**

MATHEMATIQUES APPLIQUEES de BORDEAUX

Analyse des résultats du tableau

- le critère *RA* a des valeurs fortes et assez proches les unes des autres

- le critère *RAc* est meilleur dans le cas3/3 (Tr/N en 3 classes) puis dans le cas 4/4 (Tr/nonN et 4 classes) choisi également par les critères *R* et *Rc*.

	RA	RAc
Cas 1/ 3	**0.964**	**0.655**
Cas 1/ 4	**0.952**	**0.553**
Cas 1/ 5	**0.943**	**0.561**
Cas 2/ 3	**0.953**	**0.585**
Cas 2/ 4	**0.953**	**0.631**
Cas 2/ 5	**0.942**	**0.562**
Cas 3/ 3	**0.977**	**0.782**
Cas 3/ 4	**0.943**	**0.550**
Cas 3/ 5	**0.933**	**0.498**
Cas 4/ 3	**0.963**	**0.619**
Cas 4/ 4	**0.963**	**0.714**
Cas 4/ 4	**0.942**	**0.562**

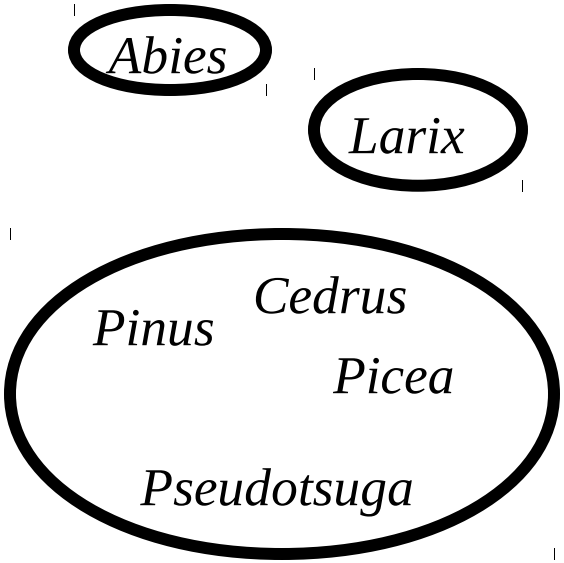
MATHEMATIQUES APPLIQUEES de BORDEAUX

Tableau croisé entre la variable *Genre* et la partition Cas4/4

	Classe1	Classe2	Classe3	Classe4	Total
Abies	0	8	0	0	8
Cedrus	1	0	2	0	3
Larix	0	0	0	6	6
Picea	4	0	3	0	7
Pinus	10	0	1	0	11
Pseudotsuga	1	0	2	0	3
Total	16	8	8	6	38

- ***Abies*** et ***Larix*** parfaitement discriminés
- ***Pinus*** presque tous dans la même classe
- Classe3 pas très « claire »

Tableau croisé entre la variable *Genre* et la partition Cas3/3



	Classe1	Classe2	Classe3	Total
Abies	**0**	**0**	**8**	**8**
Cedrus	**3**	**0**	**0**	**3**
Larix	**0**	**6**	**0**	**6**
Picea	**6**	**0**	**1**	**7**
Pinus	**10**	**0**	**1**	**11**
Pseudotsuga	**3**	**0**	**0**	**3**
Total	**22**	**6**	**10**	**38**

- ***Larix*** parfaitement discriminés, ***Abies*** presque parfaitement
- Une « grosse » classe avec les 4 autres espèces

MATHEMATIQUES APPLIQUEES de BORDEAUX

Conclusion

- Recommencer l'analyse avec plus de feuilles de chaque genre

- Critère de validation externe pour le choix de la méthode et du nombre de classes

- Découverte d'une « partition » ou de liens entre les modalités