Critère de Rand asymétrique

Application en chimie organique

M. Chavent — C. Lacomblez — B. Patouille

Mathématiques Appliquées de Bordeaux (UMR 5466)
Université Bordeaux 1, 351 cours de la libération,33405 Talence cedex
chavent@math.u-bordeaux.fr, lacomble@sm.u-bordeaux2.fr, be@sm.u-bordeaux2.fr

RÉSUMÉ. *Les critères de comparaison de deux partitions d'un même ensemble d'individus obtenues avec deux méthodes de classification différentes, sont généralement symétriques. Nous proposons une version asymétrique du critère de Rand et du critère de Rand corrigé, afin d'évaluer dans quelle mesure une partition P est "plus fine" qu'une partition Q. Ces critères seront utilisés dans le cadre d'une application en chimie organique pour choisir par validation externe une partition d'un ensemble de conifères établie à partir de la composition en acides gras des feuilles. La partition experte est quant à elle définie par la variable nominale genre.*

MOTS-CLÉS : *Validation externe, Classification de modalités, Comparaison de partitions*

1. Introduction

Un problème classique en validation externe consiste à mesurer le degré de similarité entre deux partitions d'un même ensemble de données. On a une partition *a priori* P (souvent experte) et plusieurs partitions Q obtenues par différents algorithmes de classification. Une partition sera d'autant meilleure qu'elle "ressemblera" plus à la partition experte. Deux cas de figures peuvent se présenter :

– on cherche à retrouver exactement la partition experte : P et Q doivent alors avoir le même nombre de classes et on les compare à l'aide d'un critère symétrique. Deux critères classiques pour comparer deux partitions sont le critère de Rand [RAN 71] et le critère de Rand corrigé [HUB 85]

– lorsque la partition experte est générée par une variable qualitative, on peut simplement vouloir qu'une classe de la partition obtenue contienne tous les objets d'une ou de plusieurs classes de la partition experte. P aura alors en général plus de classes que Q et il semble plus naturel d'utiliser des critères de comparaison non symétriques.

On considère donc deux partitions $P = (P_1, ..., P_i, ..., P_k)$ et $Q = (Q_1, ..., Q_j, ..., Q_l)$ d'un même ensemble Ω de n individus, le nombre k de classes de P étant supérieur au nombre l de classes de Q. Le problème est alors d'évaluer dans quelle mesure la partition P est "plus fine" que la partition Q. On considère ici qu'une partition est plus fine qu'une autre si lorsque deux éléments sont classés ensemble dans la première ils le sont également dans la seconde : $\forall i = 1, ..., k, \exists\, j$ tel que $P_i \subseteq Q_j$
On cherche ainsi à mesurer "l'inclusion" de la partition P dans la partition Q. On définira pour cela une version asymétrique et asymétrique corrigée du critère de Rand.

Ces critères seront utilisés afin de choisir par validation externe une partition d'un ensemble de conifères, la partition experte ayant été établie par les botanistes en fonction de divers critères : anatomie, morphologie, paléontologie, etc.

2. Notations et formules de passage

Les critères utilisés pour comparer deux partitions P et Q peuvent généralement être calculés soit à partir du tableau de contingence croisant ces deux partitions (Tab. 1), soit à partir du tableau accords-désaccords (Tab. 2).

	$Q_1 \cdots$	Q_j	$\cdots Q_l$	
P_1		$\vdots$		
$\vdots$				
P_i	$\cdots$	n_{ij}	$\cdots$	$n_{i.}$
$\vdots$				
P_k		$\vdots$		
		$n_{.j}$		$n_{..} = n$

Tableau 1. *Tableau de contingence*

	$\hat{\text{m}}$ classes dans Q	$\neq$ classes dans Q
$\hat{\text{m}}$ classes dans P	a	b
$\neq$ classes dans P	c	d

Tableau 2. *Tableau des accords-désaccords entre P et Q*

Le Tableau 2 indique le nombre a de paires d'individus qui sont classés ensemble pour chacune des partitions P et Q, le nombre d de paires d'individus qui sont dans deux classes différentes dans P et dans Q, le nombre b de paires qui sont classés ensemble selon Q et dans deux classes différentes selon P et inversement pour c.

On note :

$\left(\begin{array}{c} n_{ij} \\ 2 \end{array} \right) = \frac{n_{ij}(n_{ij}-1)}{2}$ le nombre de paires d'individus qui sont dans $P_i \cap Q_j$

$\left(\begin{array}{c} n \\ 2 \end{array} \right) = \frac{n(n-1)}{2}$ le nombre de paires de Ω

$\left(\begin{array}{c} n_{i.} \\ 2 \end{array} \right) = \frac{n_{i.}(n_{i.}-1)}{2}$ le nombre de paires de la classe P_i

$\left(\begin{array}{c} n_{.j} \\ 2 \end{array} \right) = \frac{n_{.j}(n_{.j}-1)}{2}$ le nombre de paires de la classe Q_j

On peut alors passer du Tableau 1 au Tableau 2 par les formules de passage suivantes [HUB85] :

$$a = \sum_{i,j} \left(\begin{array}{c} n_{ij} \\ 2 \end{array} \right)$$

$$b = \sum_{i} \left(\begin{array}{c} n_{i.} \\ 2 \end{array} \right) - \sum_{i,j} \left(\begin{array}{c} n_{ij} \\ 2 \end{array} \right)$$

$$c = \sum_{j} \left(\begin{array}{c} n_{.j} \\ 2 \end{array} \right) - \sum_{i,j} \left(\begin{array}{c} n_{ij} \\ 2 \end{array} \right)$$

$$d = \binom{n}{2} + \underbrace{\sum_{i,j}\binom{n_{ij}}{2}}_{\text{ensemble dans } P \text{ et } Q} - \underbrace{\sum_{i}\binom{n_{i.}}{2}}_{\text{ensemble dans } P} - \underbrace{\sum_{j}\binom{n_{.j}}{2}}_{\text{ensemble dans } Q}$$

Le critère de Rand et sa version asymétrique seront définis section 3 à partir des deux tableaux. En revanche, le critère de Rand corrigé et sa version asymétrique seront définis section 4 uniquement à partir du tableau de contingence.

3. Critère de Rand et sa version asymétrique

Le critère de Rand [RAN 71] que l'on note R mesure le pourcentage d'accords entre les deux partitions P et Q :

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

Ce critère est symétrique ($R(P,Q) = R(Q,P)$) et prend ses valeurs dans $[0,1]$. Si les deux partitions ont le même nombre de classes et si $\forall i, \exists j$ tel que $P_i = Q_j$, alors $R = 1$.
Dans notre problème, les partitions P et Q ne jouent pas le même rôle puisque l'on cherche à évaluer dans quelle mesure les classes de P sont incluses dans celles de Q ou encore dans quelle mesure, P est plus fine que Q. Dans ce cas, le nombre de classes de P est supérieur ou égal au nombre de classes de Q. On considère alors que les accords entre P et Q sont mesurés non seulement par a et d, mais également par c. En effet, on considère que deux éléments qui ne sont pas classés ensemble dans P peuvent l'être dans Q. Le nombre d'accords est alors $a+d+c$ et on définit le critère de Rand asymétrique, noté RA, par :

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

Ce critère est bien asymétrique ($RA(P,Q) \neq RA(Q,P)$) et prend également ses valeurs dans $[0,1]$. Si $\forall i, \exists j$ tel que $P_i \subseteq Q_j$ alors $RA = 1$.

Les critères de Rand et de Rand asymétrique s'interprètent également en terme de probabilités. Pour cela, nous reprenons le formalisme de [BEL 98] : l'ensemble fondamental E est l'ensemble de toutes les paires $\{x, y\}$ d'éléments distincts de Ω et chaque événement élémentaire $\{x, y\}$ est réalisé avec la même probabilité égale à $1/\mid E \mid$ avec $\mid E \mid = a+b+c+d$. On considère les deux événements suivants :

A = { x et y sont classés ensemble selon P }
B = { x et y sont classés ensemble selon Q }

Le critère de Rand estime donc $P(A \cap B) = \dfrac{\mid A \cap B \mid}{\mid E \mid} = \dfrac{a+d}{a+b+c+d}$. En se plaçant dans un contexte plus général de la modélisation statistique d'une règle logique, on peut dire que le critère de Rand évalue la règle

$A \Leftrightarrow B$.
Le critère de Rand asymétrique estime pour sa part $1 - P(A \cap \overline{B})$. En effet : $A \cap \overline{B}$={paires classées ensemble dans P et pas dans Q}. Donc $\mid A \cap \overline{B} \mid = b$ et

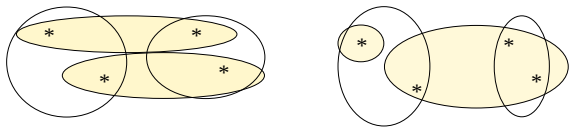
$$P(A \cap \overline{B}) = \frac{b}{a+b+c+d} = 1 - \frac{a+d+c}{a+b+c+d} = 1 - RA(P,Q)$$

Si on se place encore dans un contexte de modélisation statistique d'une règle logique, on peut dire que le critère de Rand asymétrique évalue la règle $A \Rightarrow B$. Dans [BEL 98], on trouve une autre écriture de $1 - P(A \cap \overline{B})$ proposée pour évaluer $A \Rightarrow B$, et donc une troisième écriture du critère RA:

$$1 - P(A \cap \overline{B}) = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{k} \sum_{1 \leq s < t \leq l} \overbrace{\mid P_i \cap Q_s \mid}^{n_{is}} \overbrace{\mid P_i \cap Q_t \mid}^{n_{it}}$$

4. Critère de Rand corrigé et sa version asymétrique

Comme nous l'avons vu, il est facile d'identifier les cas d'adéquation pour lesquels les critères R et RA sont égaux à 1. En revanche, il est plus difficile de déterminer les cas où R ou RA sont nuls. En effet, pour certaines configurations de partitions, ces critères ne seront jamais nuls. Par exemple, en partant de la partition P de quatre éléments en deux classes en grisé sur la Fig. 1-a), on peut trouver Q tel que $R(P,Q) = 0$. En revanche, à partir de la partition en deux classes en grisé sur la Figure 1-b), il n'existe pas de partition Q en deux classes telle que $R(P,Q) = 0$.



(a) (b)

Figure 1. *En (a) $R = 0$ et en (b) R n'est jamais nul*

Pour pallier ce problème d'échelle, [HUB 85] proposent un indice de Rand corrigé, d'espérance nulle lorsque les accords entre les deux partitions sont dus au hasard ([MIL 86]). On considère que les accords entre P et Q sont dus au hasard lorsque les deux partitions P et Q sont tirées au hasard dans l'ensemble des partititons en k et l classes respectivement, les nombres $n_{i.}$ et $n_{.j}$ d'individus par classe étant fixés.
Sous cette hypothèse nulle, la variable aléatoire N_{ij} correspondant au nombre n_{ij} d'individus dans $P_i \cap Q_j$, suit une loi hypergéométrique de paramètre $(n, n_{.j}, n_{i.})$ et [FOL 83] montrent que :

$$E\left(\sum_{i,j} \binom{n_{ij}}{2}\right) = \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}} \qquad [1]$$

On en déduit l'espérance du critère R :

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

Le critère de Rand corrigé, noté Rc, utilise la normalisation $\frac{R - E(R)}{1 - E(R)}$ et vaut donc :

$$Rc = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}}} \quad [3]$$

De la même manière, on calcule l'espérance du critère RA :

$$E(RA) = 1 + \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}^2} - \frac{\sum_i \binom{n_{i.}}{2}}{\binom{n}{2}} \quad [4]$$

Et le critère de Rand asymétrique corrigé, noté RAc, est défini par :

$$RAc = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}}}{\sum_i \binom{n_{i.}}{2} - \frac{\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}}{\binom{n}{2}}} \quad [5]$$

5. Application en botanique

Les données utilisées pour illustrer ces différents critères de comparaison sont extraites d'un fichier de données [1] concernant la composition en acides gras des feuilles d'une famille de conifères (les Pinacées) dont font partie les pins (*Pinus*), les sapins (*Abies*), les épicéas (*Picea*), les cèdres (*Cedrus*)...[MON 01]. Il s'agit d'un sous-fichier de 38 feuilles de conifères pour lesquels on connait le genre et les quantités (en pourcentage) de chacun des 24 acides gras présents dans la composition des feuilles. La variable genre est une variable nominale à 6 modalités (*Abies, Cedrus, Larix, Picea, Pinus, Pseudotsuga*) qui définit une partition experte établie par les botanistes.

Le problème est alors de construire de manière automatique à partir des 24 variables continues une partition Q des 38 feuilles. Les botanistes espèrent ainsi obtenir une partition la plus proche possible de P tout en étant moins fine. En effet, des feuilles de genres différents doivent pouvoir se retrouver dans une même classe de Q. Cette denière doit en quelque sorte définir, autant que possible une partition des modalités de la variable genre.

1. Données fournies par le Laboratoire de Biogénèse et Biotechnologie végétale (A. Badoc) et le Laboratoire de Biogenèse Membranaire (S. Mongrand)

La méthode de classification utilisée est la classification hiérarchique de Ward sur coordonnées factorielles [LEB 98]. Douze partitions Q différentes ont ainsi été obtenues avec le logiciel SPAD [LEB 96] et les options suivantes :

– Analyse en Composantes Principales (ACP) calculée soit sur les 24 variables telles quelles, soit sur les 24 variables modifiées par une transformation de Box-Cox [JOB 92] afin de corriger l'asymétrie des distributions.
– ACP normée ou non normée.

On en deduit le tableau 3 indiquant les quatre cas de figure envisagés pour des partitions en 3, 4 ou 5 classes.

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

Tableau 3. *Les différentes options de classification envisagées pour $j = 3, 4, 5$ classes*

Pour chacune des douze partitions générées, les valeurs prises par les différents critères de Rand présentés dans les sections 3 et 4 sont données dans le tableau 4 :

	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 4/5	**0.811**	**0.442**	0.942	0.562

Tableau 4. *Valeurs des différents critères pour les 12 partitions*

Ce tableau nous montre que :

– les critères de Rand et Rand corrigé sont les meilleurs pour la partition en 4 classes obtenue à partir de l'ACP non normée sur variables transformées (cas 4/4). On remarque également que les deuxièmes meilleurs résultats correspondent aux partitions en 5 classes des cas 2 et 4.
– les critères de Rand asymétriques RA et RAc sont les meilleurs pour la partition en 3 classes obtenue à partir de l'ACP normée sur variables transformées (cas 3/3). La seconde meilleure partition est celle du cas 4/4. On note ainsi que les performances du cas 4/4 sont bonnes quelque soit le critère utilisé.

L'examen de la partition correspondant au cas 4/4 (Tableau 5) montre que les individus du genre *Abies* et ceux du genre *Larix* sont parfaitement discriminés dans les classes 2 et 4 respectivement . Les *Pinus* sont tous (ou presque) dans la Classe 1, mais cette dernière contient également des *Picea* et un *Pseudotsuga*.

Dans la partition correspondant au cas 3/3 (Tableau 6), on retrouve les *Larix* et les *Abies* bien discriminés dans les classes 2 et 3 respectivement, la classe des *Abies* contenant cependant un *Pinus* et un *Picea*. La classe 1 regroupe quant à elle les deux "mauvaises" classes du cas 4/4 soit les classes 1 et 3 du tableau 5.

En conclusion, il apparaît sur cet exemple que les critères R et Rc favorisent les partitions ayant un nombre de classes proche de celui de la partition experte. Ainsi, les deux critères asymétriques RA et RAc semblent plus

	Classe 1	Classe 2	Classe 3	Classe 4	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

Tableau 5. *Tableau croisé entre la variable Genre et la partition du Cas4/4*

	Classe 1	Classe 2	Classe 3	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

Tableau 6. *Tableau croisé entre la variable Genre et la partition du Cas3/3*

appropriés lorsqu'il s'agit de comparer deux partitions ayant des nombres de classes différents, et par là-même peuvent servir ici d'indicateurs pour le choix du nombre de classes.

6. Bibliographie

[BEL 98] BEL MUFTI G., « Validation d'une classe par estimation de sa stabilité », PhD thesis, Université Paris-IX Dauphine, 1998.

[FOL 83] E.B. FOLKES, C.L. MALLOWS, « A method for comparing two Hierarchical Clusterings », *Journal of the American Statistical Association*, vol. 78, 1983, p. 553–569.

[HUB 85] HUBERT L., ARABIE P., « Comparing partitions », *Journal of Classification*, , 1985, p. 193–208.

[JOB 92] J.D. JOBSON, *Applied Multivariate Data Analysis*, vol. I:Regression and Experimental Design, Springer Verlag, 1992.

[LEB 96] LEBART L., MORINEAU A., LAMBERT T., PLEUVRET P., *SPAD version 3. Système pour l'Analyse des Données*, CISIA, Saint-Mandé, 1996.

[LEB 98] LEBART L., MORINEAU A., PIRON M., *Satistique exploratoire multidimensionnelle*, Dunod, 1998.

[MIL 86] G.W. MILLIGAN, M.C COOPER, « A study of the Comparability of External Criteria for Hierarchical Cluster Analysis », *Multivariate Behavioral Resarch*, vol. 21, 1986, p. 441–458.

[MON 01] MONGRAND S., BADOC A., CHAVENT M., LACOMBLEZ C., PATOUILLE B., CASSAGNE C., J-J. BESSOULE, « Taxonomy of gymnospermae: multivariate analysis of leaf fatty acid composition », *Phytochemistry*, , accepted March 2001.

[RAN 71] W.M. RAND, « Objective Criteria for the evaluation of Clustering Methods », *Journal of the American Statistical Association*, vol. 66, 1971, p. 846–850.