Sur des indices de comparaison de deux classifications

Genane Youness, Gilbert Saporta

Sur des indices de comparaison de deux classifications

Genane Youness
CEDRIC-CNAM
BP 114661Beyrouth, Liban
genane99@hotmail.com

Gilbert Saporta
Chaire de Statistique Appliquée et CEDRIC
CNAM
292 rue Saint Martin
75141 Paris Cedex 03
saporta@cnam.fr

RÉSUMÉ. Cet article fait suite aux travaux présentés au congrès SFC 2001, portant sur la comparaison de deux classifications d'un même ensemble d'individus. On propose d'utiliser le coefficient de corrélation vectoriel RV pour étudier la ressemblance entre deux partitions, on retrouve l'indice J de Janson et Vegelius. Le coefficient kappa de Cohen K est un autre outil pour mesurer l'accord entre les partitions, mais l'identification des classes étant nécessaire, on prend la permutation des classes qui donne le kappa maximum. On étudie la distribution d'échantillonnage de ces indices pour des paires de partitions proches car issues d'un modèle de classes latentes.

MOTS-CLÉS : kappa de Cohen, coefficient de corrélation vectoriel, indice de Janson et Vegelius, classification, classes latentes, k-means.

1. Introduction

Dans le but de comparer deux classifications provenant d'un même ensemble de données, et suite à la communication présentée au congrès SFC 2001 [SAP 01], on présente deux autres outils destinés à répondre aux questions suivantes : lors de deux enquêtes portant sur les mêmes individus, comment mesurer l'accord entre les deux classifications? Est-ce que les configurations de ces deux classifications se ressemblent ?
On présente les écritures logiques et relationnelles de l'indice dérivé du coefficient de corrélation vectorielle RV introduit par P. Robert et Y. Escoufier [ROB 76] d'où on retrouve le coefficient J introduit par S. Janson et J. Vegelius [JAN 82] ainsi que leurs distributions d'échantillonnage sous une hypothèse nulle d'absence de liaison.
Le coefficient kappa appliqué aux paires d'individus, fournit une nouvelle façon de mesurer l'accord entre deux partitions ayant le même nombre de classes, provenant d'un même échantillon. On utilise la permutation de la valeur du kappa maximal pour identifier les classes d'une partition.

2. Notation

P_1 et P_2 sont deux partitions des mêmes individus (ou deux variables qualitatives). Lorsque l'on croise deux partitions, on va s'intéresser aux paires d'individus qui restent ou ne restent pas dans les mêmes classes. On a $n(n-1)/2$ paires d'individus.

Le tableau de contingence N croisant P_1 et P_2 est de dimension pxq est caractérisé par son terme général nij= l'effectif de la case (i,j), et il est lié aux tableaux disjonctifs associées à P_1 et P_2 par la forme matricielle suivante: $N=K'_1K_2$.
Chaque partition P_k sera représentée par un tableau relationnel C^k dans l'espace des individus, de dimension nxn, dont le terme général $c^k_{ii'}$ est défini par :

$$c^k_{ii'} = \begin{cases} 1 \text{ si les deux individus } i \text{ et } i' \text{ sont dans la meme classe de la partition } P_k \\ 0 \text{ sinon} \end{cases}$$

On a $C^1=K_1K_1'$ et $C^2=K_2K_2'$
Nous posons :
n= nombre d'individus
p= nombre de classe de la partition P_1
q= nombre de classe de la partition P_2

2.1. RV ou J

Le coefficient de corrélation vectorielle RV introduit par P. Robert et Y. Escoufier [ROB 76] permet de mesurer la ressemblance entre deux études sur les mêmes observations. La procédure consiste à comparer les deux tableaux de distances nxn inter-individus ou les deux tableaux de produits scalaires en tenant compte de la possibilité d'avoir des métriques différentes pour mesurer les distances entre les individus.
Si X_1 et X_2 sont les tableaux de données numériques associées, on considère les matrices du produits scalaires entre les individus $X_iX'_i$ (ou $X_iM_iX'_i$ si l'on introduit une métrique particulière M_i à chaque matrice), notées W_i de dimension nxn on cherche à décrire les distances entre ces matrices et à comparer les tableaux entre eux. Le coefficient RV est la somme des carrées inter- covariance entre les deux tableaux X_1 et X_2 divisé par la matrice normée intra-variance.

$$RV(X_1, X_2) = \frac{\text{trace}(W_1 W_2)}{\sqrt{\text{trace}(W_1^2)\text{trace}(W_2^2)}}$$

Les travaux de A. Lazraq et R.Cleroux [LAZ 01,02] donnent la possibilité de tester des hypothèses concernant RV. Si RV est suffisamment grand, les classifications obtenues seront voisines.
Si on calcule ce coefficient pour étudier la ressemblance entre deux partitions à p et q classes respectivement, on retrouve le coefficient J introduit par S. Janson et J. Vegelius [JAN 82].

L'indice J correspond au critère de l'écart à la moyenne probabiliste CP sous l'hypothèse de l'équiprobabilité des classes pour une partition et s'écrit sous forme contingentielle ou en comparaison par paires [IDR 00] comme suit :

$$J(P_1, P_2) = \frac{pq\sum\sum n_{ij}^2 - p\sum n_{i.}^2 - q\sum n_{.j}^2 + n^2}{\sqrt{[p(p-2)\sum n_{i.}^2 + n^2][q(q-2)\sum n_{.j}^2 + n^2]}} = \frac{\sum_{i,i'}(c^1_{ii'} - \frac{1}{p})(c^2_{ii'} - \frac{1}{q})}{\sqrt{\sum_{i,i'}(c^1_{ii'} - \frac{1}{p})^2 \sum_{i,i'}(c^2_{ii'} - \frac{1}{q})^2}}$$

En effet pour deux variables nominales à p et q modalités respectivement, on passe de la forme disjonctive complète $[K_1,K_2]$ aux tableaux de comparaison par paires par $C^1=K_1K'_1$ et $C^2=K_2K_2'$, le coefficient de corrélation vectoriel s'écrit alors:

$$RV(P_1,P_2) = \frac{\text{trace}(C^1C^2)}{\sqrt{\text{trace}(C^1)^2\,\text{trace}(C^2)^2}} = \frac{\sum_{i,i'}(c^1_{ii'})(c^2_{ii'})}{\sqrt{\sum_{i,i'}(c^1_{ii'})^2 \sum_{i,i'}(c^1_{ii'})^2}}$$

Si on centre les $c^k_{ii'}$ on retrouve la forme relationnelle de l'indice J.

2.2. Le kappa de Cohen

Introduit par Cohen [COH 60], le coefficient kappa est une mesure non paramétrique d'accord entre deux variables qualitatives pour des données appariées. Il exprime la différence relative entre la proportion d'accords observés P_o et la proportion d'accords aléatoires P_e qui est la valeur espérée, sous l'hypothèse nulle d'indépendance des variables, divisée par le complément à un de l'accord aléatoire.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Si on a k modalités ou classes (dans le cas où les deux partitions auraient même nombre de classes p=q=k), la concordance observée P_o est la proportion des individus classés dans les cases diagonales de concordance du tableau de contingence, soit la somme des effectifs diagonaux divisés par la taille de l'échantillon n.

$$P_o = \frac{1}{n}\sum_{i=1}^{k} n_{ii}$$

La concordance aléatoire P_e est égale à la somme des produits des effectifs marginaux divisés par le carrée de la taille de l'échantillon.

$$P_e = \frac{1}{n^2}\sum_{i=1}^{k} n_{i.}\, n_{.i}$$

Et la forme contingentielle de l'indice de kappa est :

$$\kappa = \frac{n\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i.}n_{.i}}{n^2 - \sum_{i=1}^{k} n_{i.}n_{.i}}$$

Mais comme l'identification des classes est nécessaire pour utiliser le coefficient kappa, une valeur maximale de ce coefficient (une concordance maximale entre partitions) nous mène à identifier les classes des partitions. On renumérote les classes de l'une des deux partitions du tableau de contingence pour optimiser le kappa. On prend alors la permutation des classes qui maximise le kappa. Cette numérotation des classes n'est pas forcément celui induit par le premier axe de l'analyse factorielle des correspondances.
Par un programme en C on trouve les matrices permutées en colonnes ainsi que la valeur de kappa pour chaque permutation donc la permutation de kappa maximal. Puis on compare ce résultat au résultat trouvé par la méthode d'analyse factorielle de correspondances. On utilise donc le logiciel SPAD.
On présente les résultats trouvés pour une simulation du choix des variables, on partage à deux les variables normales indépendantes, on trouve deux partitions P_1 et P_2 à 4 classes chacune. On a 4 ! Permutations du tableau de contingence de base (P_1 en ligne, P_2 en colonne).

2.2.1 Résultats par le programme en C.

Le tableau de contingence de base est :

1	2	3	4
0.000000	128.000000	0.000000	138.000000
10.000000	0.000000	229.000000	5.000000
176.000000	0.000000	69.000000	1.000000
237.000000	0.000000	6.000000	1.000000

le tableau de contingence réordonné après calcul de la valeur de kappa maximal est :

4	3	2	1
138.000000	0.000000	128.000000	0.000000
5.000000	229.000000	0.000000	10.000000
1.000000	69.000000	0.000000	176.000000
1.000000	6.000000	0.000000	237.000000

La valeur de kappa maximal est de 0.473792, la permutation est représenté par les numéros de colonnes suivants 4,3,2,1.

2.2.2 Résultat en SPAD par la méthode CORBI

Par la méthode d'analyse factorielle de correspondance en SPAD (CORBI), on trouve les résultats suivants, en prenant en notation la variable n°1 pour présenter la partition P_1 en ligne à 4 classes et la variable n°2 pour la partition P_2 en colonne à 4 classes.

Tableau 1 : Coordonnées des fréquences actives

Libellé de la variable n°2	**Poids relatif**	**Distance à l'origine**	**Axe 1**	**Axe 2**	**Axe 3**
v21	42.30	0.99257	-0.63	0.77	0.00
v22	12.80	2.75940	1.66	0.07	-0.01
v23	30.40	1.53661	-0.56	-1.11	0.00
v24	14.50	2.41044	1.55	0.02	0.01

Par ce tableau, on peut interpréter le classement trouvé par la méthode d'analyse des correspondances, il est donc selon la permutation des numéros des colonnes suivante : 1,3,4,2. Si on cherche la valeur de kappa pour cette permutation on trouve la valeur – 0.031.
On remarque que les deux méthodes n'aboutirent pas aux même résultats, et donnent des valeurs différentes du coefficient de kappa.

2.3. Distributions d'échantillonnages

On utilise la méthodologie présentée en [SAP 02] pour étudier la distribution d'échantillonnage de ces deux indices pour des paires de partitions proches.
Rappelons ici que le but n'est pas d'étudier si les deux partitions sont indépendantes, mais si elles sont concordantes : la difficulté étant de formuler correctement l'hypothèse nulle de concordance. A partir d'une partition initiale basée sur des caractéristiques probabilistes issue d'un modèle de classes latentes, on construit deux partitions par la méthode classique des k-means (ou des nuées dynamiques). On calcule les indices ci-dessus pour les deux partitions : en itérant le procédé on obtient par simulation la distribution d'échantillonnage de RV (ou J) et de kappa.

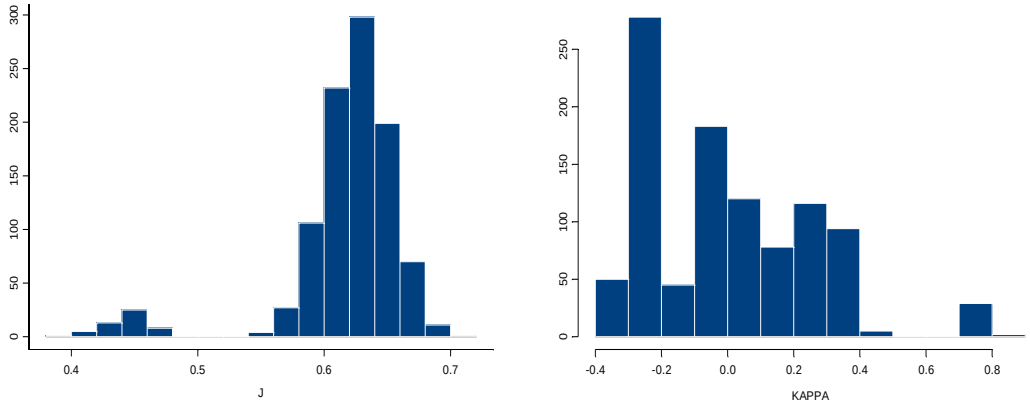

Figure 1 : La distribution du coefficient de Janson et Vegelius et celle du kappa pour pour 1000 individus et pour un choix de paramètres en 1000 itérations

Par l'une des simulations choisies on trouve que le coefficient de corrélation J varie entre 0.4 et 0.7. il y a une présence bimodale et la valeur la plus fréquente est de 0.63. La moyenne des valeurs du coefficient J est égale à 0.6172935. Pour celle du coefficient dérivé kappa de Cohen, elle varie entre –0.4 et 0.8, et de moyenne 0.0004926432. D'où l'accord entre les deux partitions.

3. Bibliographie

[COH 60] COHEN J., "A coefficient of agreement for nominal scales.*", Educ. Psychol. Meas*., vol 20, 1960, p.27-46.

[IDR 00] IDRISSI A. Contribution à l'unification de Critères d'Association pour Variables Qualitatives, Thèse de doctorat de l'Université de Paris 6, 2000.

[LAZ 02] LAZRAQ, A., CLEROUX R. *Inférence Robuste sur un indice de Redondance*, Revue de Statistique Appliquée, vol. (4), 39-54, 2002.

[JAN 82] JANSON S., VEGELIUS J. "The J - index as a measure of association for nominal scale response agreement". *Applied psychological measurement*, vol. 16, 1982, p.243-250.

[ROB 76] ROBERT P., ESCOUFIER, Y. "A unifying tool for linear multivariate statistical methods: the RV-coefficient". *Appl. Statist*., vol. 25**,** 1976, p.257-65.

[SAP 01] SAPORTA G., YOUNESS G. « Concordance entre deux partitions: quelques propositions et expériences » - *in Proceedings SFC 2001, 8èmes rencontres de la Société Francophone de Classification,* 2001, Pointe à Pitre.

[SAP 02] SAPORTA G., YOUNESS G. "Comparing two partitions: some proposals and Experiments", *Proceedings in Computational Statistics edited by Wolfgang Härdle, Physica- Verlag*, 2002, Berlin, Germany.

[SAP 03] SAPORTA G., YOUNESS G., *" Une méthodologie pour la comparaison de partitions*", Revue de Statistique Appliquée, à paraître.