Problèmes posés par la comparaison de classification dans des enquêtes différentes,

Gilbert Saporta

51ème session de l'Institut International de Statistique, Istanbul, Aout 1997

PROBLÈMES POSÉS PAR LA COMPARAISON DE CLASSIFICATIONS DANS DES ENQUÊTES DIFFÉRENTES

Gilbert SAPORTA

Conservatoire National des Arts et Métiers
292 rue Saint Martin, 75141 Paris cedex 03, France
e-mail:saporta@cnam.fr

Il est fréquent d'avoir à comparer des typologies (partitions) d'individus à l'occasion de diverses enquêtes (opinion, consommation, etc.) mais la littérature spécialisée est souvent muette sur ce point.
On proposera dans cette communication divers outils et approches destinés à répondre aux questions suivantes lors de la comparaison de deux enquêtes: peut-on affirmer que la classification n'a pas changé, que le nombre de classes est le même, que les proportions respectives des classes ont ou n'ont pas varié, que les classes s'interprétent de la même façon?
On distinguera deux cas:
a. celui où les individus sont les mêmes d'une enquête à l'autre (panel) : on peut alors avoir soit le même questionnaire, soit des questionnaires différents.
b. celui où les échantillons sont différents mais où bien sûr le questionnaire est le même.
L'existence de classes dans une population n'a pas de sens en dehors du choix d'un algorithme et d'un critère de classification: on comparera donc des classifications obtenues à l'aide de la même méthode quelle qu'elle soit. Dans nos applications il s'agira d'une méthode mixte de type « k-means » utilisable pour de grands ensembles, décrite dans Lebart, Morineau, Piron (1995).
La comparaison des résultats de deux méthodes de classification peut se traiter de la même manière que la comparaison de deux classifications du même ensemble avec les mêmes variables.
La présentation orale montrera des applications au cas b.

I. DEUX CLASSIFICATIONS SUR LES MÊMES INDIVIDUS

I.1 Questionnaires différents

I.1.1 Comparaisons de partitions sur *n* individus

Chaque partition induit une variable qualitative : l'information permettant la comparaison est alors contenue soit dans le tableau de contingence $\boldsymbol{N}$ croisant ces deux variables , soit dans les tableaux $\boldsymbol{C_1}$ et $\boldsymbol{C_2}$ *nxn* symétriques des comparaisons par paires : C^1_{ij} =1 si *i* et *j* sont dans la même classe pour la première partition, C^1_{ij} = 0 sinon.

L'utilisation des mesures habituelles d'association comme le chi-deux ne permet pas de répondre de manière adéquate à la comparaison de partitions: mesure d'écart à l'indépendance, le chi-deux n'est pas adapté au problème qui consiste à tester l'écart à une structure diagonale; l'hypothèse

d'indépendance est inintéressante ici et de toutes façons rejetée dans la plupart des cas sans que l'on puisse conclure à la concordance des partitions.

Le coefficient de Rand (1971) est mieux adapté: défini par le pourcentage du nombre de paires concordantes (c'est à dire de paires d'individus appartenant à un même classe pour les deux partitions, ou séparés pour les deux partitions), il prend ses valeurs dans l'intervalle [0;1] et vaut 1 dans le cas de partitions identiques. Il a été étudié en détail par Hubert et Arabie (1985) et Marcotorchino (1991).

La construction de tests de signification est ici plus délicate que pour l'écart à l'indépendance, car l'hypothèse nulle étant l'identité des deux partitions il faut définir un modèle probabiliste convenable. On peut se fixer par exemple un pourcentage minimal d'individus appartenant aux mêmes classes et faire un test binomial.

Ces indices et tests ne permettent pas de savoir par exemple si une partition en 5 classes obtenue dans une enquête diffère ou non de la partition en 6 classes obtenue lors d'une autre enquête.
Le plus efficace à notre avis est de soumettre à une analyse des correspondances le tableau ***N*** et de tracer dans le plan factoriel les domaines de confiance (ou de tolérance) des classes. Le nombre d'axes à considérer ici est normalement très faible, car si il y a bien correspondance entre les deux partitions, les points doivent peu s'écarter du premier axe factoriel. On pourra ainsi juger du choix du nombre de classes et de l'identité des classes des deux enquêtes.

1.1.2 Autres approches.

Toute classification repose sur les distances inter-individus: on peut donc effectuer des comparaisons en amont de la classification en comparant les deux tableaux de distances issus des deux études, ce qui suppose de définir une mètrique adaptée.
Le coefficient RV introduit par Y.Escoufier (1973) permet de mesurer la ressemblance entre deux études sur les mêmes observations:
si X_1 et X_2 sont les tableaux de données numériques associés (pour des questionnaires à variables qualitatives, on prendra les tableaux des coordonnées factorielles),

$$RV(X_1 ; X_2) = \frac{\mathrm{Trace}(W_1 W_2)}{\sqrt{\mathrm{Trace}(W_1^2)\,\mathrm{Trace}(W_2^2)}} \quad \text{où } W_i = X_i M_i X_i'$$

Les travaux de R.Cléroux et al. (1992) donnent la possibilité de tester des hypothèses concernant RV.Si RV est suffisamment grand, les classifications obtenues seront voisines.

1.2 Questionnaires identiques

En plus des propositions précédentes qui restent valables, mais ne tiennent pas compte du fait qu'il s'agit des mêmes variables, les procédures suivantes sont recommandées et s'inspirent de l'analyse discriminante:
La première étude servant d'ensemble d'apprentissage, on définit ainsi des fonctions de classement dans les groupes de la première typologie: on les applique ensuite en reclassant les individus de la deuxième enquête (qui sont les mêmes) dans la premiére : on obtient ainsi une

matrice de confusion dont l'analyse revélera les stabilités éventuelles de la typologie.Cette méthode revient à prendre une des deux enquêtes comme référentiel et à projeter l'autre dessus.

On peut aussi procéder à une étude de la différence entre les tableaux de données $D = X_1 - X_2$; plutôt que des tests de différences appariées, qui ne sont guère instructifs pour de grands échantilllons, on préferera faire une analyse typologique sur la matrice D. Si aucune structure de classification sur D n'apparait, on pourra admettre que les classifications issues de X_1 et X_2 sont semblables.

II DEUX CLASSIFICATIONS SUR DES ÉCHANTILLONS DIFFÉRENTS

Le cas se présente fréquemment lors d'enquêtes périodiques d'opinion ou de marché où le même questionnaire (ou du moins un grand nombre de questions identiques) est posé à des échantilllons différents mais de structure semblable. Il est certes théoriquement possible de tester si les échantillons sont ou ne sont pas significativement différents, mais outre que cela n'est pas facile pour des questionnaires qualitatifs, cela ne répond pas vraiment à la question.

II.1 Projection de classification

Comme en I.2, à l'aide de fonctions de classement on afffectera les unités de la deuxième enquête dans les classes de la première. Le tableau de contingence croisant classes nouvelles et classes anciennes reconstituées pourra alors être décrit et analysé par les méthodes décrites en I.1 : indice de Rand, analyse des correspondances..
On peut ainsi apprécier la stabilité des classification et tester par exemple si les poids des classes sont restés identiques.

II.2 Stabilité des interprétations

Ce qui importe est de vérifier si la signification des classes est restée la même. Pour chaque classe on identifie les variables les plus significatives à l'aide de procédures classiques de tests de comparaison de moyennes ou de pourcentage selon la nature des variables. On teste alors si, d'une enquête à l'autre, les variables significatives d'une classe ont les mêmes répartitions.

Une autre approche possible est celle basée sur les classes latentes qui sont des modèles particuliers de mélanges de distribution (cf Everitt et Hand (1981) Les parametres étant estimés par le maximum de vraisemblance (avec en général un algorithme EM) on peut alors tester l'identité de deux modèles .

II.3 Classification des variables

Comme les individus sont différents mais les variables identiques, on peut alors inverser la problématique en comparant les deux classifications de variables que l'on peut obtenir. Lorsque les classifications de variables sont semblables, cela implique de manière duale que la structuration des individus est comparable (classes de même signification, mais de poids éventuellement différents).

On est conduit à un problème au moins aussi complexe que la comparaison de classification des mêmes individus avec des variables différentes. Si on veut bâtir des tests, il faut se donner des modèles probabilistes adaptés: pour des classifications hiérarchiques il faudrait se donner une

distribution de hiérarchies sous l'hypothèse de tirages aléatoires dans une population d'individus. De tels travaux existent, basés par exemple sur des mesures de consensus, mais semblent encore difficiles à appliquer (Sokal 1988).

PERSPECTIVES

Des procédures formalisées de comparaison restent encore à définir et valider, qui reposeraient sur des modèles probabilistes d'écart à une typologie qui soient réalistes et qui tiennent compte du fait que l'appartenance à une classe comporte toujours une part d'incertitude. Les méthodes de rééchantillonnage pourraient apporter des contributions inéressantes.

RÉFÉRENCES

Escoufier Y. (1973) Le traitement des variables vectorielles, *Biometrics*, **29**, 751-760

Everitt B., Hand D.J.(1981), *Finite Mixture Distributions*, Chapman and Hall, London

Hubert L., Arabie P. (1985), Comparing partitions, *Journal of Classification*, **2**, 193-218

Lazraq A., Cléroux R., Kiers H.A.L. (1992), Mesures de liaison vectorielle et généralisation de l'analyse canonique, *Revue de Statistique Appliquée*, **39**, 23-35

Lebart L., Morineau A., Piron M. (1995), *Statistique Exploratoire Multidimensionnelle*, Dunod, Paris

Marcotorchino F., El Ayoubi N.(1991), Paradigme logique des écritures relationnelles de quelques critères d'association, *Revue de Statistique Appliquée*, **39**, 2, 25-46

Rand W.M. (1971), Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, **66**, 846-850

Sokal R.R. (1988), Unsolved problems in numerical taxonomy in *Classification and Related Methods of Data Analysis*, H.H.Bock ed., North Holland, 45-56

RÉSUMÉ

On propose diverses approches pour comparer des partitions obtenues par une méthode de classification non hiérarchique lors de deux enquêtes portant soit sur les mêmes individus, soit sur des échantillons différents mais avec le même questionnaire

SUMMARY

Several approaches are proposed for the comparison of clusters obtained by a non-hierarchical technique for two different surveys with the same units, or with two separate samples but the same questionnaire.