Détermination du nombre de classes dans les méthodes de bipartitionnement

Malika Charrad[*,***] Yves Lechevallier[**]
Gilbert Saporta[***], Mohamed Ben Ahmed[*],

[*]Ecole Nationale des Sciences de l'Informatique
malika.charrad@riadi.rnu.tn,
mohamed.benahmed@riadi.rnu.tn
[**]INRIA-Rocquencourt
yves.lechevallier@inria.fr
[***]Conservatoire National des Arts et Métiers
gilbert.saporta@cnam.fr

Résumé. Une des premières critiques que présentent certains algorithmes de partitionnement simple est la connaissance requise du nombre de classes dans la population. Le problème devient plus complexe dans le cas de bipartitionnement où il faut déterminer le nombre de classes sur les deux dimensions : les lignes et les colonnes. Nous proposons d'étendre l'application de certains indices de validation du nombre de classes proposés initialement dans le cadre de la classification simple, à la classification simultanée, en particulier à l'algorithme Croki2 de classification des tableaux de contingence. Nous proposons également un nouvel indice de validation basé sur la différentielle que nous comparons avec les indices classiques. Nous testons l'efficacité de ces méthodes sur des données générées artificiellement autour d'un nombre de classes connus sur les lignes et les colonnes.
Mots clés : Bipartitionnement, Indices de validation, Nombre de classes.

1 Introduction

Dans les méthodes de classification automatique, les auteurs définissent généralement trois types de critères de validation selon que l'on dispose ou pas d'information a priori sur les données : critère externe, critère interne et critère relatif. Nous nous intéressons dans ce papier au critère relatif qui permet de comparer deux structures de classification et choisir la structure la plus stable ou la mieux appropriée aux données en utilisant un ensemble d'indices. Ces indices de validation sont initialement proposés pour les méthodes de partitionnement simple. Dans le cadre de bipartitionnement, aucun critère n'a été proposé dans la littérature. Notre idée est d'étendre l'application de certains de ces indices aux méthodes de bipartitionnement. Pour ce faire, nous générons des données artificielles présentant une structure de biclasses sur lesquelles nous appliquons l'algorithme de bipartitionnement Croki2 proposé par Govaert (1983) pour la classification croisée des tableaux de contingence. Les indices de validation

sont appliqués aux partitions obtenues par l'algorithme Croki2 afin d'identifier les meilleurs indices capables d'estimer le bon nombre de classes sur les lignes et les colonnes.

2 Application des indices de validation aux algorithmes de bipartitionnement

Nous proposons de tester les indices suivants : l'indice de Dunn (1974), l'indice BH de Baker et Hubert (1975), l'indice DB de Davies et Bouldin (1979), l'indice CH de Calinsky et Harabsz (1974), l'indice Silhouette (S) de Rousseeuw (1987), l'indice HL de Hubert et Levin (1976), l'indice KL de Krzanowski et Lai (1988). Afin d'étendre l'utilisation de ces indices aux méthodes de bipartitionnement, nous proposons d'appliquer l'indice séparément à la partition obtenue sur les lignes et la partition obtenue sur les colonnes. Ainsi, pour chaque couple de classes nous obtenons deux valeurs de l'indice : une valeur obtenue sur la partition-ligne et une valeur obtenue sur la partition-colonne. Si la meilleure valeur de l'indice calculé sur les lignes et la meilleure valeur de l'indice calculé sur les colonnes sont obtenues pour le même couple de classes (k^*, l^*) alors nous retenons ce couple comme meilleure solution. Considérons le jeu des données JD5x4 généré autour de 5 classes artificielles sur les lignes et 4 classes artificielles sur les colonnes. La valeur maximale de l'indice BH calculé sur les lignes est obtenue pour les couples $\{(5,3),(5,4),(5,5),(5,6),(5,7)\}$. La valeur maximale de l'indice calculé sur les colonnes est obtenue pour les couples $\{(3,4),(4,4),(5,4),(6,4),(7,4)\}$. Le couple qui maximise à la fois les deux valeurs de l'indice, correspond à la bonne solution. Ainsi, la meilleure bipartition sur JD5x4 est obtenue pour 5 classes sur les lignes et 4 classes sur les colonnes. Pour certains jeux des données, en particulier les données réelles, la meilleure valeur de l'indice-ligne et la meilleure valeur de l'indice-colonne sont obtenues pour deux couples de classes différents (aucun couple en commun). Dans ce cas, la solution dépend de l'importance qu'on accorde à la partition ligne et à la partition colonne. En d'autres termes, si on privilégie la partition ligne, on choisira le couple de classes qui maximise l'indice-ligne bien qu'il ne correspond pas à la meilleure valeur de l'indice-colonne et inversement. Nous proposons alors de calculer un indice pondéré dans lequel on attribue un poids à l'indice-ligne et un poids à l'indice-colonne.

$$IndiceGlobal = \alpha * IndiceLigne + (1 - \alpha) * IndiceColonne$$

avec $\alpha \in [0, 1]$.

3 Résolution graphique/Méthode de la différentielle

Dans l'algorithme Croki2, les meilleures partitions P et Q sur les lignes et les colonnes sont celles qui maximisent le critère de χ^2 du nouveau tableau de contingence obtenu en regroupant les lignes et les colonnes suivant ces deux partitions. Or, plus le nombre de classes sur les lignes et les colonnes augmente, plus la valeur de χ^2 augmente. Cependant, à partir d'une certaine valeur du couple (k, l), par exemple (4,4) pour le jeu des données JD4x4 et (6,3) pour le jeu des données JD6x3, le χ^2 stagne ou croit très lentement. Notre objectif est alors d'identifier le couple de classes (k^*, l^*) à partir duquel le χ^2 ne croit plus ou croit plus lentement.

La fonction χ^2 qu'on notera $f(x, y)$ constitue une surface sur les axes de base x et y représentant respectivement le nombre de classes sur les lignes et le nombre de classes sur les colonnes.
Dans le cas continu, la différentielle d'ordre 2 d'une fonction réelle à deux variables $\{x, y\}$ s'écrit :

$$d^2 f = \frac{\partial^2 f}{\partial x^2}(dx)^2 + \frac{\partial^2 f}{\partial x \partial y} dxdy + \frac{\partial^2 f}{\partial y \partial x} dydx + \frac{\partial^2 f}{\partial y^2}(dy)^2$$

Afin de discrétiser ces termes, nous avons recourt aux développements limités de Taylor. La différentielle d'ordre 2 s'écrit alors :

$$d^2 f(x, y) = \frac{2f(x+h, y+h) + f(x-h, y) + f(x, y-h)}{h^2}$$
$$- \frac{f(x+h, y) + f(x, y+h) + 2f(x, y)}{h^2}$$

Pour trouver le couple de classes à partir duquel les valeurs de χ^2 ne croissent plus, nous calculons la différentielle de la fonction de χ^2 pour tous les couples de nombre de classes, en considérant $h = 1$. Le meilleur couple est celui qui correspond à la valeur la plus faible, ou la plus élevée en valeur absolue, de la différentielle d'ordre 2.

4 Comparaison des indices

Afin de comparer les indices sur des données simulées, nous générons six jeux des données JDKxL (JD3x3, JD4x4, JD5x4, JD6x3, JD6x6 et JD3x8) où K est le nombre de classes sur les lignes et L est le nombre de classes sur les colonnes. Tous les jeux des données ont la même dimension 200×100. Le tableau 1 présente les résultats de l'application des indices aux partitions obtenues sur les lignes et les colonnes en appliquant l'algorithme Croki2 à tous les jeux des données.

TAB. 1 – *Comparaison des indices sur les données simulées*

Jeu des données	**JD3x3**	**JD4x4**	**JD5x4**	**JD6x3**	**JD6x6**	**JD3x8**
Dunn	**(3,3)**	**(4,4)**	**(5,4)**	x	**(6,6)**	x
BH	**(3,3)**	**(4,4)**	**(5,4)**	**(6,3)**	**(6,6)**	(3,6)
DB	**(3,3)**	**(4,4)**	(4,4)	x	**(6,6)**	x
CH	**(3,3)**	**(4,4)**	(4,4)	x	**(6,6)**	x
Silhouette	**(3,3)**	**(4,4)**	(4,4)	x	**(6,6)**	x
HL	x	x	x	x	x	x
KL	x	x	x	x	x	x
Differentielle	**(3,3)**	**(4,4)**	**(5,4)**	**(6,3)**	**(6,6)**	(3,4)

5 Conclusion

D'après ces résultats expérimentaux, nous notons que les indices de validation Dunn, DB, CH et S sont performants dans l'identification du bon nombre de classes sur les lignes et

les colonnes lorsque les données présentent le même nombre de classes sur les lignes et les colonnes. En effet, DB, CH et S ont échoué dans l'identification du bon couple de classes dans le cas des jeux des données JD5x4, JD6x3 et JD3x8 alors qu'ils ont réussi dans le cas des jeux des données JD3x3, JD4x4 et JD6x6. L'indice BH et la différentielle ont réussi à identifier le bon couple de classes dans tous les cas sauf lorsque la différence entre le nombre de classes sur les lignes et celui sur les colonnes est grande (cas du JD3x8). Les indices HL et KL ont échoué dans tous les cas.
Ce travail peut être amélioré en testant d'autres indices proposés dans la littérature pour des méthodes classiques de classification, en particulier dans le cas où le nombre de classes sur les lignes et celui sur les colonnes sont très différents.

Références

Baker, F. B. et L. J. Hubert (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 31–38.

Calinsky, R. et J. Harabsz (1974). A dendrite method for cluster analysis. *Communications in statistics*, 1–27.

Charrad, M., Y. Lechevallier, G. Saporta, et M. B. Ahmed (2009). Block clustering for web pages categorization. *Editeurs : Corchado, Emilio ; Yin, Hujun. LNCS 5788, Intelligent Data Engineering and Automated Learning, Springer*, 260–267.

Davies, D. L. et D. W. Bouldin (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell. 1 (4)*, 224–227.

Dunn, J. C. (1974). Well separated clusters and optimal fuzzy partitions. *Journal Cybern.*, 95–104.

Govaert, G. (1983). Classification croisée. *Thèse de doctorat d'état, Paris*.

Hubert, L. et J. Levin (1976). A general statistical framework for assesing categorical clustering in free recall. *Psychological Bulletin*, 1072–1080.

Krzanowski, W. et Y. Lai (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics 44*, 23–34.

Rousseeuw, P. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 53–65.

Summary

Some clustering algorithms require the knowledge of the number of clusters in the population. The problem becomes more complex with biclustering algorithms where it is necessary to determine the number of clusters on both dimensions: rows and columns. We propose to extend the application of some indices used for clustering validation to biclustering validation, mainly with Croki2 algorithm. We also propose a new validation index based on the total differential that we compare with conventional indices. We test the efficiency of these methods on data artificially generated around a number of clusters known on the rows and columns.
Key words : Block clustering, Validity indices, Number of clusters.