Sélection de mesures de proximité pour une analyse des correspondances topologique

Rafik Abdesselam

Sélection de mesures de proximité pour une analyse des correspondances topologique

Rafik Abdesselam

Laboratoire COACTIS-ISH, UFR de Sciences Economiques et de Gestion
Université de Lyon, Lumière Lyon 2,
16, quai Claude Bernard 69365 Lyon cedex 07
rafik.abdesselam@univ-lyon2.fr

Résumé. L'approche proposée consiste à comparer puis à classer des mesures de proximité dans un contexte topologique afin de sélectionner la meilleure mesure en vue d'effectuer une analyse des correspondances topologique. Les mesures de similarité jouent un rôle important dans de nombreux domaines de l'analyse des données. Les résultats de toute opération de classification ou de classement d'objets dépendent fortement de la mesure de proximité utilisée. Basées sur la notion de graphes de voisinage, certaines mesures de proximité sont plus ou moins équivalentes. Un indice d'quivalence topologique entre deux mesures de proximité est défini dans le cadre de l'association entre deux variables qualitatives. Un exemple sur données réelles illustre cette mthode.

Mots-clés. Mesure de proximité, graphe de voisinage, matrice d'adjacence, équivalence topologique, association et indépendance.

Abstract. The proposed approach consists of comparing and classifying proximity measures in a topological context in order to select the best measure for performing a topological correspondence analysis. Similarity measures play an important role in many domains of data analysis. The results of any clustering or classification operation of objects depend strongly on the measure of proximity used. Based on the notion of neighborhood graphs, some proximity measures are more or less equivalent. A topological independence index between two proximity measures is defined in the framework of the association between two qualitative variables. An example on real data illustrates this approach.

Keywords. Proximity measure, neighborhood graph, adjacency matrix, topological equivalence, association and independence.

1 Introduction

Le choix d'une mesure de proximité est un problème important. La comparaison d'objets, de situations ou d'idées est une tâche essentielle pour évaluer une situation, classer des préférences ou structurer un ensemble d'éléments. Pour ce faire, nous utilisons des mesures de proximité pour mettre en évidence les similarités ou les dissimilarités entre objets. Nous savons pertinemment que le résultat dépend de la mesure utilisée. Alors, laquelle est la plus utile ? Sont-elles équivalentes ? Comment identifier celle qui est la plus appropriée

pour résumer la liaison entre deux variables qualitatives ? Selon la mesure choisie, les résultats de cette problématique d'analyse des correspondances topologique changent.

Nous nous intéressons ici à la caractérisation d'un indice d'indépendance topologique entre deux variables qualitatives. Plus cet indice est grand, plus les deux variables sont indépendantes selon la mesure de proximité choisie. Plusieurs études sur l'équivalence topologique de mesures de proximité ont été proposées (Zighed *et al.* (2012)), mais aucune dans un objectif d'association entre variables qualitatives. Nous avons considéré et comparé 22 mesures de proximité conventionnelles de la littérature pour des données qualitatives (Warrens (2008)).

Soit x et y deux variables qualitatives à p et q modalités-groupes respectivement décrivant un ensemble de n individus-objets. On utilise les notations suivantes :
- $X_{(n,p)}$ la matrice associée aux p indicatrices $\{x^j; j = 1, p\}$ de x,
- $Y_{(n,q)}$ la matrice associée aux q indicatrices $\{y^k; k = 1, q\}$ de y,
- $K_{(p,q)} = {}^tX\, Y$ le tableau de contingence,
- $Z_{(n,r)} = [\, X \,|\, Y \,] = [\, z^1 = x^1, \cdots, z^j = x^j, \cdots, z^p = x^p \,|\, z^{p+1} = y^1, \cdots, z^k = y^k, \cdots, z^r = y^q \,]$ le tableau disjonctif complet, juxtaposition des tableaux binaires X et Y, à n lignes-objets et $r = p + q$ colonnes-modalités,
- $M_{B(r,r)} = {}^tZ\, Z = \left(\begin{array}{c|c} {}^tX\, X & {}^tX\, Y \\ \hline {}^tY\, X & {}^tY\, Y \end{array}\right) = \left(\begin{array}{c|c} {}^tX\, X & K \\ \hline {}^tK & {}^tY\, Y \end{array}\right)$ le tableau de Burt associé,
- $W_{(r,r)} = Diag[M_B] = \left(\begin{array}{c|c} {}^tX\, X & 0 \\ \hline 0 & {}^tY\, Y \end{array}\right) = \left(\begin{array}{c|c} W_p & 0 \\ \hline 0 & W_q \end{array}\right)$ matrice diagonale des effectifs des $r = p + q$ modalités de x et y,
- I_r la matrice identité d'ordre r,
- $U_{(r,r)} = 1_r\ {}^t1_r$ la matrice dont tous les éléments sont égaux à 1, 1_r et 1_n désignent respectivement le vecteur d'ordre r et d'ordre n de composantes égales à 1.

Les matrices de dissimilarité associées aux mesures de proximité sont calculées à partir du tableau de contingence K. Le temps de calcul est ainsi considérablement réduit.

$A_{(r,r)} = (a_{jk}) = M_B$ dont l'élément, $a_{jk} = |Z^j \cap Z^k| = \sum_{i=1}^n z_i^j z_i^k$ correspond au nombre d'attributs communs aux deux points z^j et z^k,

$B_{(r,r)} = (b_{jk}) = {}^tZ\,(1_n\,{}^t1_r - Z) = {}^tZ\,1_n\,{}^t1_r\ -\ {}^tZ\,Z = W\,1_r\,{}^t1_r - A = W\,U - A$

dont l'élément, $b_{jk} = |Z^j - Z^k| = |Z^j \cap \overline{Z^k}| = \sum_{i=1}^n z_i^j(1 - z_i^k)$ est le nombre d'attributs présents dans z^j mais pas dans z^k,

$C_{(r,r)} = (c_{jk}) = {}^t(1_n\,{}^t1_r\ -\ Z)\,Z = {}^t(1_n\,{}^t1_r)\,Z - {}^tZ\,Z = 1_r\,{}^t1_n\,Z\ -\ {}^tZ\,Z = UW - A$

dont l'élément, $c_{jk} = |Z^k - Z^j| = |Z^k \cap \overline{Z^j}| = \sum_{i=1}^n z_i^k(1 - z_i^j)$ est le nombre d'attributs présents dans z^k mais pas dans z^j.

$D_{(r,r)} = (d_{jk}) = {}^t(1_n\,{}^t1_r\ -\ Z)\,(1_n\,{}^t1_r\ -\ Z) = nU - (A + B + C)$

dont l'élément, $d_{jk} = |\overline{Z^j} \cap \overline{Z^k}| = \sum_{i=1}^n (1 - z_i^j)(1 - z_i^k)$ est le nombre d'attributs qui ne sont présents ni dans z^j ni dans z^k.

$Z^j = \{i / z_i^j = 1\}$ et $Z^k = \{i / z_i^k = 1\}$ étant les ensembles d'attributs présents respectivement dans les données du point-modalité z^j et z^k et $|.|$ désigne le cardinal de l'ensemble. Les quatre quantités sont liées : $\forall j = 1, p\,;\, \forall k = 1, q\ a_{jk} + b_{jk} + c_{jk} + d_{jk} = n$.

2 Indice d'indépendance topologique

L'équivalence topologique repose sur la notion de graphe topologique que l'on désigne également par graphe de voisinage. Deux mesures de proximité sont équivalentes si les graphes topologiques induits sur l'ensemble des objets restent identiques. Mesurer la ressemblance entre mesures de proximité revient à comparer les graphes de voisinage et à mesurer leur ressemblance.

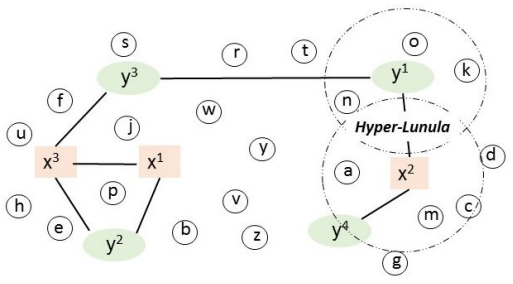
Soit l'ensemble $E = \{z^1 = x^1, \ldots, z^p = x^p, z^{p+1} = y^1, \ldots, z^r = y^q\}$ à $r = |E|$ modalités dans $\{0,1\}^n$, associé aux variables x et y. On peut, à l'aide d'une mesure de proximité u, définir une relation de voisinage V_u qui sera une relation binaire sur $E \times E$.

Pour une mesure de proximité donnée u, on peut construire un graphe de voisinage sur l'ensemble des objets-modalités, où, les sommets sont les modalités et les arêtes sont définies par une relation de voisinage.

Il existe de nombreuses définitions pour construire la relation binaire de voisinage. Par exemple, on peut recourir aux Graphes des Voisins Relatifs (GVR), (Toussaint, (1980)), dont les couples de points voisins (z^j, z^k) vérifient la propriété GVR suivante :

$$\begin{cases} V_u(z^j, z^k) = 1 & if\ u(z^j, z^k) \leq \max[u(z^j, z^l), u(z^l, z^k)]\ ;\ \forall z^j, z^k, z^l \in E,\ z^l \neq z^j\ et\ z^k \\ V_u(z^j, z^k) = 0 & ailleurs \end{cases}$$

Pour toute mesure de proximité donnée u, on peut lui associer une matrice dite d'adjacence V_u binaire et symétrique d'ordre $r = p + q$. La Figure 1 illustre un ensemble de n objets-individus autour de sept modalités associées à deux variables qualitatives x et y respectivement à trois et à quatre modalités. Par exemple, $V_u(z^2 = x^2, z^4 = y^1) = 1$ cela signifie sur le plan géométrique, que l'hyper-Lunule (intersection des deux hypersphères centrées sur les deux points-modalités x^2 et y^1) est vide.



V_u	$z^1 = x^1$	$z^2 = x^2$	$z^3 = x^3$	$z^4 = y^1$	$z^5 = y^2$	$z^6 = y^3$	$z^7 = y^4$
$z^1 = x^1$	1	0	1	0	1	0	0
$z^2 = x^2$	0	1	0	1	0	0	1
$z^3 = x^3$	1	0	1	0	1	1	0
$z^4 = y^1$	0	1	0	1	0	1	0
$z^5 = y^2$	1	0	1	0	1	0	0
$z^6 = y^3$	0	0	1	1	0	1	0
$z^7 = y^4$	0	1	0	0	0	0	1

Figure 1: Exemple GVR à sept modalités - Matrice d'adjacence associée

- **Comparaison et sélection de mesures de proximité**

Pour mesurer l'équivalence topologique entre deux mesures de proximité u_i et u_j, nous proposons de tester si les matrices d'adjacence associées V_{u_i} et V_{u_j} sont différentes ou non. L'indice d'équivalence topological entre deux matrices d'adjacence est mesuré par la propriété de concordance suivante :

$$S(V_{u_i}, V_{u_j}) = \frac{\sum_{k=1}^{r}\sum_{l=1}^{r}\delta_{kl}(z^k,z^l)}{r^2} \quad \text{avec} \quad \delta_{kl}(z^k, z^l) = \begin{cases} 1 & \text{si} \quad \mathrm{V}_{u_i}(z^k, z^l) = V_{u_j}(z^k, z^l) \\ 0 & \text{ailleurs.} \end{cases}$$

La mesure de similarité $S(V_{u_i}, V_{u_j}) = 1$ signifie que les deux matrices d'adjacence sont identiques et par conséquent, la structure topologique induite par les deux mesures est la même. Dans ce cas, on parle d'équivalence topologique parfaite entre les deux mesures de proximité. La valeur $S(V_{u_i}, V_{u_j}) = 0$ signifie que la topologie a totalement changé.

On note $V_{u*} = I_r$ la matrice d'adjacence, d'indépendance parfaite entre les deux variables, associée à une mesure de proximité dite de référence, inconnue, notée u_*. On peut alors établir l'Indice d'Indépendance Topologique $IIT_i = S(V_{u_i}, V_{u*})$ entre les deux variables, en mesurant la similarité entre les matrices d'adjacence V_{u_i} et V_{u*}. Plus cet indice est grand, plus les variables sont indépendantes selon la mesure de proximité u_i choisie. Pour visualiser les mesures de proximité, on peut par exemple, appliquer une classification ascendante hiérarchique (CAH) sur les facteurs significatifs de l'analyse en composantes principales (ACP) du tableau des dissimilarités $[D]_{ij} = 1 - S(V_{u_i}, V_{u_j})_{i,j=1,22}$. De plus, pour déterminer la classe de mesures de proximité la plus proche de la mesure de référence u_*, cette dernière sera considérée comme élément illustratif dans les analyses, en projetant *a posteriori* le vecteur de dissimilarités $[D]_{*i} = 1 - S(V_{u*}, V_{u_i})_{i=1,22}$.

- **Comparaison statistique de l'équivalence topologique**

Soient V_{u_i} et V_{u_j} les matrices d'adjacence associées à deux mesures de proximité u_i et u_j. Pour comparer le degré d'équivalence topologique entre deux mesures de proximité, nous testons si les matrices d'adjacence associées sont statistiquement différentes ou pas, en utilisant le test non paramétrique de Kappa (Cohen, (1960)). Ces matrices binaires et symétriques d'ordre r, sont dépliées selon deux vecteurs de composantes appariées, formées des $\frac{r(r-1)}{2}$ valeurs supérieures (ou inférieures) de la diagonale. Le degré d'équivalence topologique entre les deux mesures u_i et u_j est évalué à partir du coefficient de concordance de Kappa, calculé sur le tableau 2×2 de contingence formé par les deux vecteurs :

$$\kappa = \kappa(V_{u_i}, V_{u_j}) = \frac{P_o - P_e}{1 - P_e}$$

$$\text{avec} \begin{cases} \mathrm{P}_o = \frac{2}{r(r+1)}\sum_{k=0}^{1} n_{kk} & \text{la proportion de concordance observée,} \\ \mathrm{P}_e = \frac{4}{r^2(r+1)2}\sum_{k=0}^{1} n_{k.}n_{.k} & \text{la proportion de concordance attendue sous l'hypothèse d'indépendance.} \end{cases}$$

Il y a une parfaite indépendance d'accord ou de concordance sous l'hypothèse nulle $H_0 : \kappa = 0$. La concordance est d'autant plus élevée que sa valeur tend vers $+1$, parfaite ou maximale si $\kappa = 1$ et une parfaite discordance lorsque $\kappa = -1$.

- **Exemple d'application**

Pour illustrer cette approche, on considère les données de l'INSEE[1] présentées dans le tableau de contingence de la Table 1 et qui concernent les créations d'entreprises en France - 2016. Dans un contexte métrique, l'hypothèse nulle du test d'indépendance du khi-deux est clairement rejetée avec un risque d'erreur $\alpha \leq 5\%$. Il existe donc une forte liaison entre le type d'entreprise et le secteur d'activité de la création.

[1]Institut National de la Statistique et des Etudes Economiques

Secteur d'activité \ Type d'entreprise	Société	Entreprise individuelle classique	Micro-Entrepreneur	Total
Industrie	8,6	7,7	8,3	24,6
Construction	26,5	18,6	16,5	61,6
Commerce, Transport, Hébergement et Restauration	64	48,7	48,7	161,5
Information et Communication	11,1	2,1	14,5	27,6
Activités Financières et d'Assurance	12,6	1,3	2	15,8
Activités Immobilières	11,3	5,1	2,5	18,9
Activités Scientiques et Techniques Spcialisées	27,6	11,9	51	90,6
Education et Santé	6,5	26,4	36,4	69,4
Activités de Service	20,6	20,6	42,9	84
Total	188,8	142,4	222,8	554

Table 1: Table de contingence - Création d'entreprises (en milliers)

Lettre Mesure	IIT (%)	$\widehat{\kappa}(V_{u_i}, V_{u*})$	p-value
A Jaccard	62.50	0.308	$< .0001$
A Dice, Czekanowski	62.50	0.308	$< .0001$
A Kulczynski	62.50	0.308	$< .0001$
A Driver, Kroeber and Ochiai	62.50	0.308	$< .0001$
A Sokal and Sneath 2	62.50	0.308	$< .0001$
A Braun and Blanquet	62.50	0.308	$< .0001$
A Simpson	62.50	0.308	$< .0001$
A Russel and Rao	62.50	0.308	$< .0001$
A Sokal and Sneath 5	62.50	0.308	$< .0001$
A Baroni, Urbani and Buser	62.50	0.308	$< .0001$
A Q-Yule	62.50	0.308	$< .0001$
A Y-Yule	62.50	0.308	$< .0001$
B Sokal and Sneath 4	70.83	0.397	$< .0001$
C Pearson	73.61	0.432	$< .0001$
D Michael	73.61	0.432	$< .0001$
E Simple Matching	84.72	0.606	$< .0001$
E Rogers and Tanimoto	84.72	0.606	$< .0001$
E Hamann	84.72	0.606	$< .0001$
E BC	84.72	0.606	$< .0001$
E Sokal and Sneath 3	84.72	0.606	$< .0001$
E Gower and Legendre	84.72	0.606	$< .0001$
E Sokal and Sneath 1	84.72	0.606	$< .0001$

Table 2: Indice d'Indépendance Topologique & Test de Kappa

Dans un contexte topologique, les principaux résultats de l'approche proposée sont présentés dans les tableaux et graphiques suivants. Ils permettent de visualiser les mesures de proximité proches les unes des autres dans un contexte d'indépendance entre le type d'entreprise et le secteur d'activité de la création.

La Table 2 présente par ordre croissant les $IIT_i = S(V_{u_i}, V_{u*})$ ainsi que les valeurs de la statistique de test de Kappa entre la mesure de référence u_* et chacune des 22 mesures de proximité considérées. Plus cet indice est grand, plus on se rapproche de la position d'indépendance et plus l'hypothèse nulle sera rejetée. Les 22 mesures considérées rejettent toutes l'hypothèse nulle $H_0 : \kappa = 0$ (pas de concordance - indépendance), elles concluent toutes à l'existence d'un lien entre le type d'entreprise et le secteur d'activité.

On a également calculé les similarités $S(V_{u_i}, V_{u_j})$ et les coefficients de Kappa $\widehat{\kappa}(V_{u_i}, V_{u_j})$ des $C_{22}^2 = 231$ couples de mesures de proximité. Tous les tests statistiques sont significatifs avec un risque d'erreur $\alpha \leq 5\%$. Les similarités par paire diffèrent quelque peu, certaines sont plus proches que d'autres. Dans la Table 2, les mesures avec la même lettre sont en parfaite équivalence topologique $S(V_{u_i}, V_{u_j}) = 1$ avec une concordance parfaite $\widehat{\kappa}(V_{u_i}, V_{u_j}) = 1$, elles sont donc identiques.

Le dendrogramme de la Figure 2 et la Table 3 résument les principaux résultats de la partition retenue en quatre classes homogènes de mesures de proximité. La mesure de référence u_* associée à une indépendance parfaite entre le type d'entreprise et le secteur d'activité, est affectée à la classe 1. La liaison entre ces deux variables qualitatives sera plus faible si l'on choisit une mesure de proximité parmi celles de la classe 1, elle sera plus forte si l'on choisit une mesure de proximité parmi celles de la classe 4.

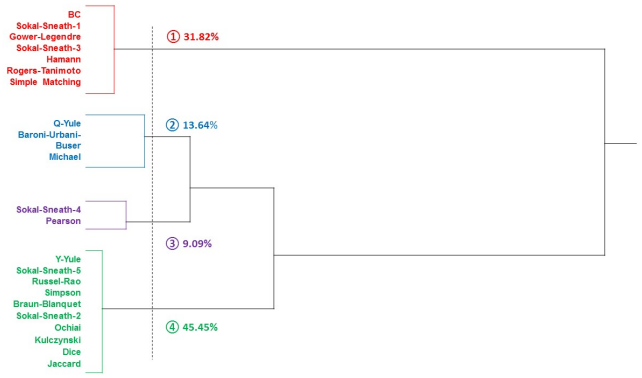


Figure 2: Arbre hiérarchique des mesures de proximité

Numéro de la classe Effectif	Classe 1 7	Classe 2 3	Classe 3 2	Classe 4 10
Mesures de proximité	$u_{Simple-Matching}$ $u_{Rogers-Tanimoto}$ u_{Hamann} u_{BC} $u_{Sokal-Sneath-3}$ $u_{Gower-Legendre}$ $u_{Sokal-Sneath-1}$	$u_{Michael}$ $u_{Baroni-Urbani-Buser}$ u_{Q-Yule}	$u_{Pearson}$ $u_{Sokal-Sneath-4}$	$u_{Jaccard}$ u_{Dice} $u_{Kulczynski}$ u_{Ochiai} $u_{Sokal-Sneath-2}$ $u_{Braun-Blanquet}$ $u_{Simpson}$ $u_{Russel-Rao}$ $u_{Sokal-Sneath-5}$ u_{Y-Yule}
Mesure de référence	$u*$			

Table 3: Affectation de la mesure de référence

3 Conclusion et perspectives

Ce travail propose une approche de sélection de mesures de proximité dans un contexte d'indépendance topologique entre deux variables qualitatives dont le but est d'établir une Analyse des Correspondances Topologique. Il serait intéressant d'étendre cette approche au cas de plus de deux variables, l'Analyse des Correspondances Multiple Topologique.

Bibliographie

[1] Cohen, J. (1960), A coefficient of agreement for nominal scales. *Educ Psychol Meas*, Vol 20, 27–46.
[2] Toussaint, G. T. (1980), The relative neighbourhood graph of a finite planar set. *In Pattern recognition*, 12, 4, 261–268.
[3] Warrens, M. J. (2008), Bounds of resemblance measures for binary (presence/absence) variables. In Journal of Classification, *Springer*, 25, 2, 195–208.
[4] Zighed, D., Abdesselam, R., and Hadgu, A. (2012), Topological comparisons of proximity measures. *16th PAKDD 2012 Conference*, Part I, LNAI 7301, Springer, 379–391.