Equivalence topologique entre mesures de proximité

Djamel Abdelkader Zighed*, Rafik Abdesselam**, Ahmed Bounekkar***

Laboratoire ERIC, Université Lumière Lyon 2
5 Avenue Pierre Mendès-France, 69676 Bron Cedex, France

*abdelkader.zighed@univ-lyon2.fr,
http://eric.univ-lyon2.fr/~zighed
**rafik.abdesselam@univ-lyon2.fr
http://eric.univ-lyon2.fr/~rabdesselam/fr/
***ahmed.bounekkar@univ-lyon1.fr
http://eric.univ-lyon2.fr/

Résumé. Le choix d'une mesure de proximité entre objets a un impact direct sur les résultats de toute opération de classification supervisée ou pas, de comparaison, d'évaluation ou de structuration d'un ensemble d'objets. Pour un problème donné, l'utilisateur est amené à choisir une parmi les nombreuses mesures de proximité existantes. Or, selon la notion d'équivalence topologique choisie, certaines sont plus ou moins équivalentes. Dans cet article, nous proposons une nouvelle approche pour choisir puis comparer les mesures de proximité dans un but de discrimination. A cet effet, nous introduisons un nouveau concept baptisé équivalence topologique. Ce dernier fait appel à la structure de voisinage local. Nous proposons alors de définir l'équivalence topologique entre deux mesures de proximité à travers la structure topologique induite par chaque mesure. Nous illustrons le principe de ce choix et de cette comparaison sur un exemple simple pour une quinzaine de mesures de proximités de la littérature.

1 Introduction

Comparer des objets, des situations ou des idées sont des tâches essentielles pour identifier quelque chose, évaluer une situation, structurer un ensemble d'éléments matériels ou abstraits etc. En un mot pour comprendre et agir, il faut savoir comparer. Cette comparaison, que le cerveau accomplit naturellement, doit cependant être explicitée si l'on veut la faire accomplir à une machine. Pour cela, on fait appel aux mesures de proximité.

Les mesures de proximité sont caractérisées par des propriétés mathématiques précises. Sont-elles, pour autant, toutes équivalentes ? Peuvent-elles être utilisées dans la pratique de manière indifférenciée ? Autrement dit, est-ce que, par exemple, la mesure de proximité entre individus plongés dans un espace multidimensionnel comme R^p, influence ou pas le résultat des opérations comme la classification en groupes ou la recherche des k-plus-proches voisins ?

Cette problématique est importante dans les applications pratiques. Par exemple, pour la recherche d'information dans une base de données ou sur internet. En soumettant une requête à

un moteur de recherche, de manière rapide, celui-ci nous retourne une liste de réponses classées selon leur degré de pertinence par rapport à la requête. Ce degré de pertinence peut alors être perçu comme une mesure de dissimilarité/similarité entre la requête et les objets disponibles dans la base. Est-ce que la façon dont on mesure la similarité ou la dissimilarité entre objets affecte le résultat d'une requête ? Si oui, comment décider de quelle mesure de similarité ou de dissimilarité il faut se servir. Il en est de même quand dans de nombreux domaines on souhaite réaliser un regroupement des individus en classes. La manière de mesurer la distance impacte directement la composition des groupes obtenus.

Une mesure de proximité peut être définie de manière intuitive, par exemple, comme le fait Lin (1998) et selon les hypothèses retenues ou les axiomes requis, cela débouche sur des mesures ayant des propriétés diverses et variées. Le terme proximité recouvre des significations telles que la similarité, la ressemblance, la dissimilarité, la dissemblance, etc. On trouve dans la littérature des dizaines de mesures différentes, notamment si on prend en compte la diversité des types de données (binaires, quantitatifs, qualitatifs, flou...). Dès lors, le choix de la mesure de proximité reste posé. Certes, le contexte d'application, les connaissances a priori, le type de données etc., peuvent aider à identifier les mesures idoines. Par exemple, si les objets à comparer sont décrits par des vecteurs booléens, on peut se limiter à une catégorie de mesures spécifiquement dédiées. Néanmoins, comment faire quand le nombre de mesures candidates reste grand ? Si toutes les mesures étaient équivalentes, il suffirait d'en prendre une au hasard. Pour faire face à ce problème de comparaison et de choix entre mesures de proximités , trois approches sont utilisées.

1. Par agrégation de mesures : il s'agit déviter de choisir une mesure particulière. Par exemple, Richter (1992) utilise, sur un même jeu de données, plusieurs mesures de proximité et agrège ensuite, arithmétiquement, les résultats partiels de chacune en une valeur unique. Le résultat final, peut être perçu comme une synthèse des différents points de vues exprimés par chaque mesure de proximité. Cette approche, évite ainsi de traiter la question de la comparaison qui reste cependant un problème en soi.
2. Par évaluation empirique : de nombreux travaux exposent des méthodologies pour comparer les performances des différentes mesures de proximité. Pour cela il est fait appel soit à des benchmarks comme dans Liu et al. , Strehl et al. (2000) dont les résultats attendus sont connus préalablement, soit à des critères jugés pertinents pour l'utilisateur et qui permettent, in fine, d'identifier la mesure de proximité la plus appropriée. On peut citer quelques travaux dans cette catégorie comme Noreault et al. (1980), Malerba et al. (2002), Spertus et al. (2005).
3. Par comparaison : l'objectif des travaux qui se situent dans cette catégorie vise à comparer les mesures de proximité entre elles. Par exemple, on vérifie si elles ont des propriétés communes Clarke et al. (2006), Lerman (1967) ou si l'une peut s'exprimer en fonction de l'autre Zhang et Srihari (2003), Batagelj et Bren (1995) ou simplement si elles fournissent le même résultat sur une opération de classification Fagin et al. (2003), etc. Dans ce cas précis, les mesures de proximité peuvent alors être catégorisées selon leur degré de ressemblance. L'utilisateur peut ainsi identifier les mesures qui sont équivalentes de celles qui le sont le moins Lesot et al. (2009), Bouchon-Meunier et al. (1996).

Le travail que nous proposons dans ce papier se situe dans la troisième catégorie qui vise à comparer les mesures de proximité entre elles afin de détecter celles qui sont identiques de

celles qui le sont moins. Il s'agit en fait de les regrouper en classes selon leurs similitudes. Pour comparer deux mesures de proximité, l'approche consiste, jusque là, à comparer les valeurs des matrices de proximité induites Batagelj et Bren (1995), Bouchon-Meunier et al. (1996), et, le cas échéant, à établir un lien fonctionnel explicite quand les mesures sont équivalentes. Pour comparer deux mesures de proximité, Lerman (1967) s'intéresse aux préordres induits par les deux mesures de proximité et évalue leur degré de ressemblance par la concordance entre les préordres induits sur l'ensemble des couples d'objets. D'autres auteurs, Schneider et Borlund (2007b) évaluent l'équivalence entre deux mesures par un test statistique entre les matrices de proximité. Les indicateurs numériques issus de ces comparaisons croisées servent alors à catégoriser les mesures. L'idée commune à ces travaux de comparaison s'appuie sur un postulat qui dit que deux mesures de proximité sont d'autant plus proches que les préordres induits sur les couples d'objets ne changent pas. On donnera plus loin des définitions plus précises. Dans ce papier, nous proposons une autre définition. Pour cela, on va s'intéresser à la structure de voisinage des objets que l'on appellera la structure topologique induite par la mesure de proximité. Si la structure de voisinage entre objets, induite par une mesure de proximité u_i ne change pas par rapport à celle d'une autre mesure de proximité u_j, cela signifie que les ressemblances locales entre individus n'ont pas changées. Dans ce cas, on dira que les mesures de proximité u_i et u_j sont en équivalence topologique. On pourra ainsi calculer une mesure d'équivalence topologique entre les couples de mesures de proximité et effectuer ensuite une classification sur les mesures de proximité. Nous allons définir cette nouvelle approche et montrer les premiers liens que nous avons identifiés entre elle et celle basée sur la préordonnance. A ce jour, nous n'avons pas trouvé de publications qui abordent le problème sous le même angle que nous. Ce papier est organisé comme suit. Dans la section 2, nous allons décrire de manière plus précise le cadre théorique dans lequel nous nous plaçons et nous rappelons les définitions classiques et notamment l'approche basée sur la préordonnance induite. Dans la section 3, nous introduisons notre approche d'équivalence topologique. Dans la section 4, on effectuera quelques évaluations en comparaison avec les anciennes approches et on tentera de mettre en évidence les liens entre les deux démarches. Les ouvertures qu'offre notre approche seront détaillées en conclusion.

2 Comparaison de mesures de proximité

Une mesure de proximité entre objets peut être définie selon d'une part les propriétés mathématiques requises et, d'autre part, l'espace de description des objets à comparer. Dans cet article, nous allons nous restreindre aux mesures de proximité construite sur R^p. Nous verrons dans la partie conclusion et perspectives que notre approche peut être étendue à n'importe quel type de mesure de proximité, qu'elle soit binaire Batagelj et Bren (1995), Lerman (1967), Warrens (2008), Lesot et al. (2009), floue Zwick et al. (1987), Bouchon-Meunier et al. (1996), symbolique Malerba et al. (2002), Malerba et al. (2001), etc.

2.1 Les mesures de proximité et leurs propriétés

On considère un échantillon de n individus $x, y, \ldots$ plongés dans un espace à p dimensions. Les individus sont décrits par des variables continues : $x = (x_1, \ldots, x_p)$. Une mesure de proximité u entre deux points-individus x et y de R^p est définie comme suit :

$$\begin{array}{rcl} u : R^p \times R^p & \longmapsto & R \\ (x, y) & \longmapsto & u(x, y) \end{array}$$

avec les propriétés suivantes, $\forall (x, y) \in R^p \times R^p$:

P1 : $u(x, y) = u(y, x)$ P2 : $u(x, x) \geq u(x, y)$ P2' : $u(x, x) \leq u(x, y)$ P3 : $\exists \alpha \in R \ u(x, x) = \alpha$

Une mesure de proximité u vérifiant les propriétés P1 et P2 est une mesure de ressemblance. Si elle vérifie les propriétés P1 et P2' c'est une mesure de dissemblance. Il est facile de montrer que toute mesure de ressemblance r peut être transformée en une mesure de dissemblance d comme suit : $r(x, y) = -\ d(x, y)$.

On peut également définir une mesure de proximité δ : $\delta(x, y) = u(x, y) - \alpha$ qui vérifie les propriétés suivantes, $\forall (x, y) \in R^p \times R^p$:

T1 : $\delta(x, y) \geq 0$ T2 : $\delta(x, x) = 0$ T3 : $\delta(x, x) \leq \delta(x, y)$

Une mesure de proximité qui vérifie les propriétés T1, T2 et T3 est une mesure de dissimilarité. On peut également citer d'autres propriétés comme :

T4 : $\delta(x, y) = 0 \Rightarrow \forall z \in R^p \ \ \delta(x, z) = \delta(y, z)$ T5 : $\delta(x, y) = 0 \Rightarrow x = y$
T6 : $\delta(x, y) \leq \delta(x, z) + \delta(z, y)$ T7 : $\delta(x, y) \leq \max(\delta(x, z), \delta(z, y))$
T8 : $\delta(x, y) + \delta(z, t) \leq \max((\delta(x, z) + \delta(y, t)), (\delta(x, t) + \delta(y, z)))$

On trouve dans Batagelj et Bren (1992) quelques relations entre ces inégalités :

T7*(Inég. Ultramétrique)* $\Rightarrow$ T6*(Inég. Triangulaire)* $\Leftarrow$ T8*(Inég. de Buneman)*

Une mesure de dissimilarité qui vérifie les propriétés T5 et T6 est une distance.

Il faut souligner le fait que ces propriétés ne sont pas spécifiques aux mesures de proximité construites sur R^p. Nous donnons en annexe Tableau 1 quelques mesures de proximité classiques définies sur R^p. Il convient de noter, que certaines mesures supposent que les valeurs x_i soient toutes positives. C'est ce que nous gardons pour nos expérimentations.

2.2 Comparaison de deux indices de proximité

Il est facile de constater que sur un même jeu de données, deux mesures de proximité u_i et u_j conduisent généralement à des matrices de proximité différentes. Peut-on dire que ces deux mesures de proximité sont différentes ? De nombreux articles ont été consacré à cette question. On peut trouver dans Lerman (1967) une proposition qui consiste à dire que deux mesures de promiximité u_i et u_j sont équivalentes dès lors que le préordre induit par chacune des mesures sur tous les couples d'objets sont identiques. d'où la définition suivante.

Equivalence en préordonnance Soient n objets $x, y, z...$ de R^p quelconques et deux mesures de proximité u_i et u_j sur ces objets. Si pour tout quadruplé (x, y, z, t), on a :
$u_i(x, y) \leq u_i(z, t) \Rightarrow u_j(x, y) \leq u_j(z, t)$ alors les deux mesures u_i et u_j sont considérées comme équivalentes.

Cette définition a été ensuite reprise dans de nombreux papiers Batagelj et Bren (1995), Bouchon-Meunier et al. (1996), Lesot et al. (2009) et Schneider et Borlund (2007a) mais ce dernier, ne cite pas Lerman (1967). Cette définition débouche sur un théorème intéressant dont on peut trouver la démonstration dans Batagelj et Bren (1995).

Théorème 1 Soient deux mesures de proximité u_i et u_j, s'il existe une fonction f strictement monotone telle que pour tout couple d'objets (x, y) on a $u_i(x, y) = f(u_j(x, y))$ alors u_i et u_j induisent des préordres identiques et par conséquent, elles sont équivalentes : $u_i \equiv u_j$. La réciproque étant également vraie, i.e. deux mesures de proximité dont l'une est fonction de l'autre induisent le même préordre et sont, par conséquent, équivalentes.

On peut alors proposer d'utiliser un indice de discordance entre préordres induits comme mesure de proximité entre deux mesures u_i et u_j. A cet effet, on peut, à l'instar de Rifqi et al. (2003) utiliser le tau de Kendall généralisé qui repose sur la mesure de concordance des rangs. Les rangs des $n(n-1)$ paires de valeurs de proximité entre x et y selon u_i sont comparés à ceux selon u_j. On note $R_i(x, y)$ et $R_j(x, y)$ les rangs respectifs de $u_i(x, y)$ et $u_j(x, y)$.

$$K_{u_i,u_j} = \frac{2}{n(n-1)} \sum_x \sum_{y \neq x} \delta_{ij}(x, y) \qquad \text{avec} \qquad \delta_{ij} = \begin{cases} 0 & \text{si } R_i(x, y) = R_j(x, y) \\ 1 & \text{sinon} \end{cases}$$

Cette définition montre ainsi que l'équivalence ne repose pas sur les valeurs numériques des deux matrices mais sur les préordres induits sur les couples de points. La comparaison entre indices de proximité a été étudiée par Schneider et Borlund (2007a,b) sous un angle statistique. Les auteurs proposent une approche empirique qui vise à comparer les matrices de proximité obtenues par chaque mesure de proximité sur les couples d'objets. Ils proposent ensuite de tester si les matrices sont statistiquement différentes ou pas en utilisant le test de Mantel, Mantel (1967). Le critère utilisé par ces auteurs est le coefficient des rangs de Spearman :

$$\rho_s = 1 - \frac{6 \sum_x \sum_{y \neq x} (R_i(x, y) - R_j(x, y))^2}{n(n^2 - 1)}$$

Les mêmes auteurs proposent de traiter la comparaison des préordre induits par les mesures de proximité dans le cadre de l'analyse de Procruste. Ces techniques visant à comparer directement des matrices de proximité ont été développées pour des domaines appliqués comme l'écologie, les sciences sociales, la géographie, la psychologie et l'anthropologie. Dans ce travail, nous ne discutons pas du choix de la mesure de comparaison des matrices de proximité. Nous nous contentons d'utiliser l'expression présentée plus haut. Nous re-précisons que notre objectif n'est pas de comparer des matrices de proximité ni les préordres induits mais de proposer une autre notion qui est l'équivalence topologique que nous comparons à l'équivalence préordinale en essayant d'identifier les liens entre les deux approches.

3 Equivalence topologique

L'équivalence topologique repose en fait sur la notion de graphe topologique que l'on désigne également sous le nom de graphe de voisinage. L'idée de base est en fait assez simple : 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. Nous allons tout d'abord définir de manière plus précise ce qu'est un graphe topologique et comment le construire. Nous proposons ensuite une mesure de proximité entre graphes topologiques qui servira à comparer les mesures de proximité dans la section d'après.

3.1 Graphe topologique

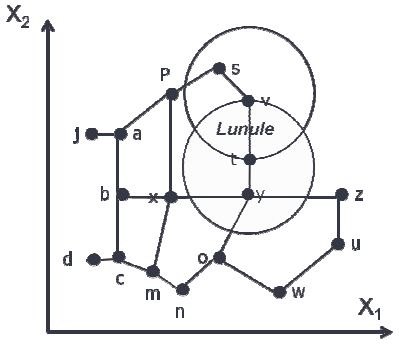
Sur un ensemble de point $x, y, z, \ldots$ de R^p, on peut, au moyen d'une mesure de proximité u_i définir une relation de voisinage V_u qui sera une relation binaire sur $E \times E$. Pour simplifier la compréhension mais sans nuire à la généralité du propos, considérons un ensemble d'objet $E = \{x, y, z, \ldots\}$ de $n = |E|$ objets plongés dans R^p. Il existe de nombreuses possibilités pour construire une relation binaire de voisinage.

Par exemple, on peut construire l'arbre de longueur minimale sur $(E \times E)$ et dire que deux objets x et y vérifient la propriété de voisinage selon l'Arbre de Longueur Minimale (ALM), Kim et Lee (2003), s'ils sont reliés par une arête directe. Dans ce cas, $V_u(x, y) = 1$ sinon $V_u(x, y) = 0$. Où V_u est la matrice d'adjacence associée au graphe ALM, formée de 0 et de 1.

On peut utiliser de nombreuses définitions pour construire la relation binaire de voisinage. Par exemple, on peut recourir aux Graphes des Voisins Relatifs (GVR), Toussaint (1980); Preparata et Shamos (1985), dont tous les couples de points voisins vérifient la propriété suivante :

$$u_E(x, y) \leq \max(u_E(x, z), u_E(y, z)) \; ; \; \forall z \in E - \{x, y\}$$

ce qui signifie, sur le plan géométrique, que l'hyper-Lunule (intersection des deux hypersphères centrées sur les deux points) est vide. La figure 1 montre un résultat dans R^2. Dans ce cas, $u_E(x, y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$ est la distance euclidienne.



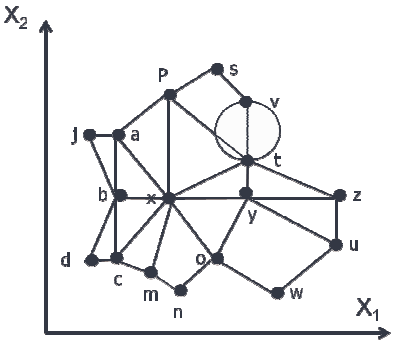
	...	x	y	z	t	u	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
x	...	1	1	0	0	0	...
y	...	1	1	1	1	0	...
z	...	0	1	1	0	1	...
t	...	0	1	0	1	0	...
u	...	0	0	1	0	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...

FIG. 1 – *Exemple de GVR pour un ensemble de points dans R^2 et la matrice d'adjacence associée.*

De manière analogue, on peut utiliser le Graphe de Gabriel (GG), Park et al. (2006), dont tous les couples de points vérifient :

$$u_E(x, y) \leq min(\sqrt{u_E^2(x, z) + u_E^2(y, z)}) \; ; \; \forall z \in E - \{x, y\}$$

Autrement dit, l'hypersphère de diamètre $u_E(x, y)$ est vide. Ce qui donnerait, sur l'exemple dans R^2, le graphe de voisinage de la figure 2.



	...	x	y	z	t	u	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
x	...	1	1	0	1	0	...
y	...	1	1	1	1	0	...
z	...	0	1	1	1	1	...
t	...	1	1	1	1	0	...
u	...	0	0	1	0	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...

FIG. 2 – *Exemple de GG pour un ensemble de points dans R^2 et la matrice d'adjacence associée.*

3.2 Proximité entre graphes topologiques

Pour fixer les idées, considérons deux mesures de proximité u_i et u_j prises parmi celles que nous avons recensées en annexe tableau 1. Prenons par exemple, la distance euclidienne $u_E(x, y)$ et la distance de Mahalanobis $u_{Mah}(x, y)$, et soient $D_E(E \times E)$ et $D_{Mah}(E \times E)$ les tableaux des distances associés.

Pour une propriété de voisinage donnée, chacune de ces deux distances engendre une structure topologique sur les objets E. Une telle structure est parfaitement décrite par sa matrice d'adjacence. On notera V_E et V_{Mah} les deux matrices d'adjacence associées aux deux structures toplogiques. Pour mesurer le degré de ressemblance entre graphes, il suffit de compter le nombre de discordances entre les deux matrices d'adjacence. La matrice étant symétrique, on peut alors calculer cette quantité par :

$$D(V_E, V_{Mah}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}}{N^2} \qquad \text{avec} \qquad \delta_{ij} = \begin{cases} 0 & \text{si } V_E(i,j) = V_{Mah}(i,j) \\ 1 & \text{sinon} \end{cases}$$

D est la mesure de disimilarité qui varie dans l'intervalle $[0, 1]$. La valeur 0 signifie que les deux matrices d'adjacence sont identiques et par conséquent, la structure topologique induite par les deux mesures de proximité est la même. Dans ce cas, on parle d'équivalence topologique entre les deux mesures de proximité. La valeur 1 signifie que la topologie a totalement changé, autrement dit, aucun couple de voisins dans la structure topologique induite par la première mesure de proximité, n'est resté voisin dans la structure topologique induite par la seconde mesure et vice versa. D s'interprète également comme le pourcentage de désaccord entre des tableaux d'adjacence.

Grâce à cette mesure de proximité, nous allons enfin pouvoir comparer les mesures de proximité et les classer selon leur degré de ressemblance. Nous verrons que les résultats obtenus sur ces classifications sont différents. En effet, une équivalence topologique n'implique pas une équivalence en préordonance. En revanche, une équivalence en préordonance entraîne une équivalence topologique.

4 Classification des mesures de proximité

Nous nous limitons dans ce travail à la classification des mesures de proximité dans R^p. Ce travail peut être étendu à toutes les autres mesures dès lors qu'on est capable de construire une structure topologique sur les objets. Nous considérons un jeu de données relativement simple, celui des Iris de Fisher. Pour construire la structure topologique, nous utilisons la propriété du graphe des voisins relatifs Toussaint (1980).

Le tableau de dissimilarité entre les 13 mesures de proximité est donné en annexe tableau 2. L'application d'un algorithme de construction d'une hiérarchie de partition selon le critère de ward Ward Jr (1963) permet d'obtenir le dendogramme suivant.

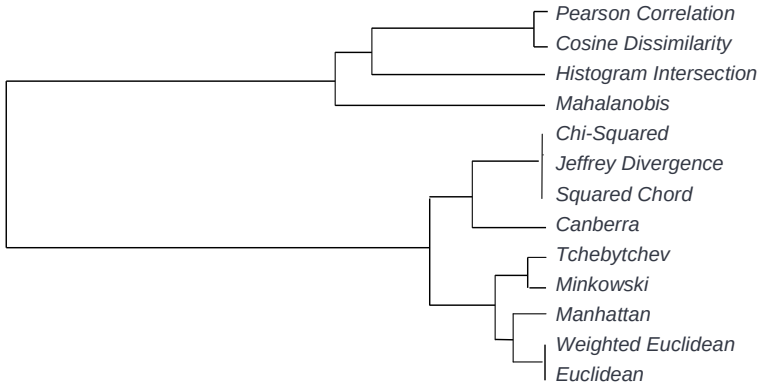


FIG. 3 – *Arbre hiérarchique - Topologie.*

4.1 Comparaison

Si nous comparons les mêmes mesures selon le critère de préordonnance, nous obtenons la matrice de dissimilarité donnée en annexe tableau 2 et le dendogramme suivant.

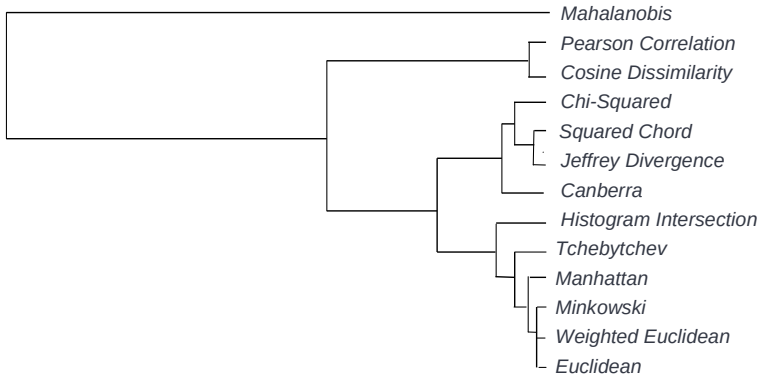


FIG. 4 – *Arbre hiérarchique - Préordonnance.*

On constate que les résultats de la classification diffèrent selon que l'on compare les mesures de proximité au moyen de l'équivalence de préordonnance ou de l'équivalence topologique.

Nous allons maintenant montrer quelques résultats plus généraux. Du théorème 1 d'équivalence en préordonnance, on en déduit la propriété suivante.

Propriété Soient f une fonction strictement monotone de R^+ dans R^+, u_i et u_j deux mesures de proximité telles que : $u_i(x,y) \rightarrow f(u_i(x,y)) = u_j(x,y)$ alors :
$u_i(x,y) \leq max(u_i(x,z)\,,\, u_i(y,z)) \Leftrightarrow u_j(x,y) \leq max(u_j(x,z)\,,\, u_j(y,z))$.

Démontration Supposons que : $max(u_i(x,z)\,,\, u_i(y,z)) = u_i(x,z)$,
d'après le théorème 1 d'équivalence en préordonnance,

$$u_i(x,y) \leq u_i(x,z) \Rightarrow f(u_i(x,y)) \leq f(u_i(x,z)),$$

de plus,
$$u_i(y,z) \leq u_i(x,z) \Rightarrow f(u_i(y,z)) \leq f(u_i(x,z))$$
$$\Rightarrow f(u_i(x,z)) \leq max(f(u_i(x,z)), f(u_i(y,z))),$$

d'où le résultat, $u_j(x,y) \leq max(u_j(x,z), u_j(y,z))$.

L'implication réciproque est vraie, vu que f est continue et strictement monotone alors, son application réciproque f^{-1} est continue et de même sens de variation que f.

On peut ainsi dire, dans le cas où f est strictement monotone, que si le préordre est conservé alors la topologie est conservée et inversement. Cette propriété nous amène à énnoncer le théorème suivant.

Théorème 2 - Equivalence en topologie Soient deux mesures de proximité u_i et u_j, s'il existe une fonction f strictement monotone telle que pour tout couple d'objets (x,y) on a $u_i(x,y) = f\,(u_j(x,y))$ alors u_i et u_j induisent des graphes topologiques identiques et par conséquent, elles sont équivalentes : $u_i \equiv u_j$. La réciproque étant également vraie, i.e. deux mesures de proximité dont l'une est fonction de l'autre induisent la même topologie et sont, par conséquent, équivalentes.

La proposition ci-dessous montre que l'équivalence en préordonnance de deux mesures de proximité u_i et $u_j = f(u_i)$ implique nécessairement l'équivalence en topologie, quelque soit la fonction f.

Proposition Dans le cadre des structures topologiques induites par le graphe des voisins relatifs, si deux mesures de proximité u_i et u_j sont équivalentes en préordonnance, alors elles sont en équivalence topologique.

Démontration Si $u_i \equiv u_j$ (équivalence en préordonnance) alors,

$$u_i(x,y) \leq u_i(z,t) \Rightarrow u_j(x,y) \leq u_j(z,t) \qquad \forall x,y,z,t \in R^p$$

on a, en particulier pour $t = x = y$ et $z \neq t$,

$$\begin{cases} \mathtt{u}_i(x,y) \leq u_i(z,x) \Rightarrow u_j(x,y) \leq u_j(z,x) \\ \mathtt{u}_i(x,y) \leq u_i(z,y) \Rightarrow u_j(x,y) \leq u_j(z,y) \end{cases}$$

on en déduit,

$$u_i(x,y) \leq max(u_i(z,x), u_i(z,y)) \Rightarrow u_j(x,y) \leq max(u_j(z,x), u_j(z,y))$$

en utilisant la propriété P1 de symétrie,

$$u_i(x,y) \leq max(u_i(x,z), u_i(y,z)) \Rightarrow u_j(x,y) \leq max(u_j(x,z), u_j(y,z))$$

d'où, $u_i \equiv u_j$ (équivalence topologique).

5 Conclusion et perspectives

Le choix d'une mesure de proximité est très subjectif, il est souvent fondé sur des habitudes ou sur des critères tels que l'interprétation a posteriori des résultats. Ce travail propose une nouvelle approche d'équivalence entre mesures de proximité. Cette approche que nous appelons topologique est basée sur la notion de graphe de voisinage induit par la mesure de de proximité. D'un point de vue pratique, dans ce papier, les mesures que nous avons comparées sont toutes construites sur des données quantitatives. Mais ce travail peut parfaitement s'étendre aux autres en choisissant la bonne structure topologique adaptée.

Nous envisageons d'étendre ce travail à d'autres structures topologiques et d'utiliser un critère de comparaison, autre que les techniques de classification, afin de valider le degré d'équivalence entre deux mesures de proximité. Par exemple, un critère basé sur un test non paramétrique (la corrélation de rang de Spearman, le tau de concordance de rang de Kendall, ou encore le test de Mantel, etc.). L'application d'un test par permutations, sur les matrices d'adjacence associées à ces mesures, vont permettre de donner une signification statistique entre les deux matrices de ressemblance et de valider ou pas l'équivalence topologique c'est-à-dire, si vraiment elles induisent ou pas la même structure de voisinage sur les objets.

Batagelj et Bren (1995) ; Malerba et al. (2001) ; Rifqi et al. (2003)

Références

Batagelj, V. et M. Bren (1992). Comparing resemblance measures. Technical report, Proc. International Meeting on Distance Analysis (DISTANCIA'92).

Batagelj, V. et M. Bren (1995). Comparing resemblance measures. *Journal of classification 12*, 73–90.

Bouchon-Meunier, B., M. Rifqi, et S. Bothorel (1996). Towards general measures of comparison of objects. *Fuzzy sets and systems 84*(2), 143–153.

Clarke, K., P. Somerfield, et M. Chapman (2006). On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted bray-curtis coefficient for denuded assemblages. *Journal of Experimental Marine Biology & Ecology 330*(1), 55–80.

Fagin, R., R. Kumar, et D. Sivakumar (2003). Comparing top k lists. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 36. Society for Industrial and Applied Mathematics.

Kim, J. et S. Lee (2003). Tail bound for the minimal spanning tree of a complete graph. *Statistics Probability Letters 64*(4), 425–430.

Lerman, I. (1967). *Indice de similarité et préordonnance associée, Ordres*. Travaux du séminaire sur les ordres totaux finis, Aix-en-Provence.

Lesot, M.-J., M. Rifqi, et H. Benhadda (2009). Similarity measures for binary and numerical data: a survey. *IJKESDP 1*(1), 63–84.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Volume 296304. Citeseer.

Liu, H., D. Song, S. Ruger, R. Hu, et V. Uren. Comparing dissimilarity measures for content-based image retrieval. *Information Retrieval Technology*, 44–50.

Malerba, D., F. Esposito, V. Gioviale, et V. Tamma (2001). Comparing dissimilarity measures for symbolic data analysis. *Proceedings of Exchange of Technology and Know-how and New Techniques and Technologies for Statistics 1*, 473–481.

Malerba, D., F. Esposito, et M. Monopoli (2002). Comparing dissimilarity measures for probabilistic symbolic objects. *Series Management Information Systems 6*, 31–40.

Mantel, N. (1967). A technique of disease clustering and a generalized regression approach. *Cancer Research 27*, 209–220.

Noreault, T., M. McGill, et M. Koll (1980). A performance evaluation of similarity measures, document term weighting schemes and representations in a boolean environment. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pp. 76. Butterworth & Co.

Park, J., H. Shin, et B. Choi (2006). Elliptic gabriel graph for finding neighbors in a point set and its application to normal vector estimation. *Computer-Aided Design 38*(6), 619–626.

Preparata, F. et M. Shamos (1985). *Computational geometry: an introduction*. Springer.

Richter, M. (1992). Classification and learning of similarity measures. *Proceedings der Jahrestagung der Gesellschaft f ur Klassifikation, Studies in Classification, Data Analysis and Knowledge Organisation. Springer Verlag*.

Rifqi, M., M. Detyniecki, et B. Bouchon-Meunier (2003). Discrimination power of measures of resemblance. *IFSA'03*.

Schneider, J. et P. Borlund (2007a). Matrix comparison, part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal American Society for Information Science and Technology 58*(11), 1586–1595.

Schneider, J. et P. Borlund (2007b). Matrix comparison, part 2: Measuring the resemblance between proximity measures or ordination results by use of the mantel and procrustes statistics. *Journal American Society for Information Science and Technology 58*(11), 1596–1609.

Spertus, E., M. Sahami, et O. Buyukkokten (2005). Evaluating similarity measures: a large-scale study in the orkut social network. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 684. ACM.

Strehl, A., J. Ghosh, et R. Mooney (2000). Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pp. 58–64.

Toussaint, G. (1980). The relative neighbourhood graph of a finite planar set. *Pattern recognition 12*(4), 261–268.

Ward Jr, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*(301), 236–244.

Warrens, M. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification 25*(2), 195–208.

Zhang, B. et S. Srihari (2003). Properties of binary vector dissimilarity measures. In *Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing*. Citeseer.

Zwick, R., E. Carlstein, et D. Budescu (1987). Measures of similarity among fuzzy concepts: A comparative analysis. *INT. J. APPROX. REASON. 1*(2), 221–242.

Annexe

Mesure	Formule
u_1 : Euclidean	$u_E(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
u_2 : Mahalanobis	$u_{Mah}(x,y) = \sqrt{(x-y)^t \sum^{-1}(x-y)}$
u_3 : Manhattan (City-block)	$u_{Man}(x,y) = \sum_{i=1}^{p} \lvert x_i - y_i \rvert$
u_4 : Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{i=1}^{p} \lvert x_i - y_i \rvert^\gamma)^{\frac{1}{\gamma}}$
u_5 : Tchebytchev	$u_{Tch}(x,y) = \max_{1 \leq i \leq p} \lvert x_i - y_i \rvert$
u_6 : Cosine Dissimilarity	$u_{Cos}(x,y) = 1 - \frac{<x,y>}{\lVert x \rVert \lVert y \rVert}$
u_7 : Canberra	$u_{Can}(x,y) = \sum_{i=1}^{p} \frac{\lvert x_i - y_i \rvert}{\lvert x_i \rvert + \lvert y_i \rvert}$
u_8 : Squared Chord	$u_{SC}(x,y) = \sum_{i=1}^{p}(\sqrt{x_i} - \sqrt{y_i})^2$
u_9 : Weighted Euclidean	$u_{E_w}(x,y) = \sqrt{\sum_{i=1}^{p} \alpha_i (x_i - y_i)^2}$
u_{10} : Chi-square	$u_{\chi^2}(x,y) = \sum_{i=1}^{p} \frac{(x_i - m_i)^2}{m_i}$
u_{11} : Jeffrey Divergence	$u_{JD}(x,y) = \sum_{i=1}^{p}(x_i \log \frac{x_i}{m_i} + y_i \log \frac{y_i}{m_i})$
u_{12} : Histogram Intersection Measure	$u_{HIM}(x,y) = 1 - \frac{\sum_{i=1}^{p}(\min(x_i, y_i))}{\sum_{j=1}^{p} y_j}$
u_{13} : Pearson's Correlation Coefficient	$u_\rho(x,y) = 1 - \lvert \rho(x,y) \rvert$

TABLE 1 – *Quelques mesures de proximité.*

Où, p est la dimension de l'espace, $x = (x_i)_{i=1,\ldots,p}$ et $y = (y_i)_{i=1,\ldots,p}$ deux points de R^p, $(\alpha_i)_{i=1,\ldots,p} \geq 0$, $\sum^{-1}$ l'inverse de la matrice des variances covariances, $\gamma > 0$, $m_i = \frac{x_i + y_i}{2}$ et $\rho(x,y)$ désigne le coefficient de corrélation linéaire de Bravais-Pearson.

D	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{13}
u_1	**1**	.776	.973	.988	.967	.869	.890	.942	1	.947	.945	.926	.863
u_2	.876	**1**	.773	.774	.752	.701	.707	.737	.776	.739	.738	.742	.703
u_3	.964	.840	**1**	.964	.940	.855	.882	.930	.973	.933	.932	.924	.848
u_4	.964	.876	.947	**1**	.967	.871	.892	.946	.988	.950	.949	.925	.866
u_5	.947	.858	.929	.964	**1**	.865	.887	.940	.957	.942	.942	.914	.860
u_6	.858	.858	.840	.840	.858	**1**	.893	.898	.869	.899	.899	.830	.957
u_7	.911	.840	.929	.893	.911	.822	**1**	.943	.890	.940	.942	.874	.868
u_8	.947	.840	.947	.929	.947	.858	.947	**1**	.942	.995	.998	.913	.884
u_9	1	.876	.964	.964	.947	.858	.911	.947	**1**	.947	.945	.926	.863
u_{10}	.947	.840	.947	.929	.947	.858	.947	1	.947	**1**	.998	.912	.885
u_{11}	.947	.840	.947	.929	.947	.858	.947	1	.947	1	**1**	.914	.884
u_{12}	.884	.813	.884	.867	.902	.884	.884	.920	.884	.920	.920	**1**	.825
u_{13}	.867	.849	.831	.867	.867	.973	.796	.849	.867	.849	.849	.876	**1**

TABLE 2 – *Tableaux des dissimilarités - Topologie (ligne) & Préordonnance (colonne).*

Les éléments situés au-dessus de la diagonale principale correspondent aux dissimilarités en préordonnance et ceux au-dessous correspondent aux dissimilarités en topologie.

Summary

The choice of a proximity measure between objects has a direct impact on the results of any operation of supervised or unsupervised classification, comparison, evaluation or structuring a set of objects. For a given problem, the user is prompted to choose one among the many existing proximity measures. However, according to the notion of topological equivalence chosen, some are more or less equivalent. In this paper, we propose a new approach to select and compare the proximity measures for the purpose of discrimination. In a context of discrimination, we introduce a new concept of topological equivalence. This approach exploits the concept of local neighborhood and believes that two proximity measures are equivalent if they induce the same neighborhood structure on the objects. We illustrate the principle of this selection and comparison on a simple example for about fifteen proximity measures of the literature.