Choix d'une mesure de proximité discriminante dans un contexte topologique

Fatima-Zahra Aazi*, Rafik Abdesselam**

*Laboratoires ERIC & $LM2CE$
Universités Lumière Lyon 2, France & Hassan 1er, Settat, Maroc
5, avenue Pierre Mendès-France, 69676 Bron Cedex, France
aazi.zf@gmail.com
**COACTIS-ISH, Université de Lyon, Lumière Lyon 2,
14/16, avenue Berthelot, 69363 Lyon Cedex 07, France
rafik.abdesselam@univ-lyon2.fr
http://eric.univ-lyon2.fr/~rabdesselam/fr/

Résumé. Les résultats de toute opération de classification ou de classement d'objets dépendent fortement de la mesure de proximité choisie. L'utilisateur est amené à choisir une mesure 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 de comparaison et de classement de mesures de proximité, dans une structure topologique et dans un objectif de discrimination. Le concept d'équivalence topologique 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 dans un contexte de discrimination. Nous proposons également un critère pour choisir la "meilleure" mesure adaptée aux données considérées, parmi quelques mesures de proximité les plus utilisées dans le cadre de données quantitatives. Le choix de la "meilleure" mesure de proximité discriminante peut être vérifié *a posteriori* par une méthode d'apprentissage supervisée de type SVM, analyse discriminante ou encore régression Logistique, appliquée dans un contexte topologique.

Le principe de l'approche proposée est illustré à partir d'un exemple de données quantitatives réelles avec huit mesures de proximité classiques de la littérature. Des expérimentations ont permis d'évaluer la performance de cette approche topologique de discrimination en terme de taille et/ou de dimension des données considérées et de sélection de la "meilleur" mesure de proximité discriminante.

1 Introduction

La comparaison d'objets, de situations ou d'idées sont des tâches essentielles pour évaluer une situation, pour classer des préférences ou encore pour structurer un ensemble d'éléments

matériels ou abstraits, etc. En un mot pour comprendre et agir, il faut savoir comparer. Ces comparaisons que le cerveau accomplit naturellement, doivent cependant être explicitées si l'on veut les faire accomplir à une machine. Pour cela, on fait appel aux mesures de proximité. Une mesure de proximité est une fonction qui mesure la ressemblance ou la dissemblance entre deux objets d'un ensemble. Ces mesures de proximité ont des propriétés mathématiques et des axiomes précis. Mais est-ce que ces mesures sont toutes équivalentes ? Peuvent-elles être utilisées dans la pratique de manière indifférenciée ? Produiront-elles les mêmes bases d'apprentissage qui serviront comme entrée pour l'estimation de la classe d'appartenance d'un nouvel objet. Si nous savons que la réponse est non, alors comment pouvoir décider laquelle utiliser ? Certes, le contexte de l'étude ainsi que le type de données considérées peuvent aider à sélectionner quelques mesures de proximités, mais laquelle choisir parmi cette sélection ?

On retrouve cette problèmatique dans le cadre d'une classification supervisée ou d'une discrimination. L'affectation ou le classement d'un objet anonyme à une classe dépend en partie de la base d'apprentissage utilisée. Selon la mesure de proximité choisie, cette base d'apprentissage change et par conséquent le résultat du classement aussi.

On s'intéresse ici au degré d'équivalence topologique de discrimination de ces mesures de proximité. Plusieurs études d'équivalence topologique de mesures de proximité ont été proposées Batagelj et Bren (1992, 1995); Rifqi et al. (2003); Lesot et al. (2009); Zighed et al. (2012) mais pas dans un objectif de discrimination. Cet article met donc l'accent sur la façon de construction la matrice d'adjacence induite par une mesure de proximité, tout en tenant compte des classes d'appartenance des objets, connues *a priori*, en juxtaposant des matrices d'adjacence intra-groupe et inter-groupes Abdesselam (2014). Un critère de sélection de la "meilleure" mesure est proposé. On vérifie en effet *a posteriori* qu'elle est bien une bonne mesure discriminante en utilisant la méthode des SVM multi-classes.

Mesure	Formule
Euclidean	$u_E(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
Mahalanobis	$u_{Mah}(x,y) = \sqrt{(x-y)^t \sum^{-1}(x-y)}$
Manhattan	$u_{Man}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert$
Tchebychev	$u_{Tch}(x,y) = \max_{1\leq j\leq p} \lvert x_j - y_j \rvert$
Cosine dissimilarity	$u_{Cos}(x,y) = 1 - \frac{\sum_{j=1}^{p} x_j y_j}{\sqrt{\sum_{j=1}^{p} x_j^2}\sqrt{\sum_{j=1}^{p} y_j^2}} = 1 - \frac{<x,y>}{\|x\|\|y\|}$
Normalized Euclidean	$u_{NE}(x,y) = \sqrt{\sum_{j=1}^{p}(\frac{x_j - y_j}{\sigma_j})^2}$
Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{j=1}^{p} \lvert x_j - y_j \rvert^\gamma)^{\frac{1}{\gamma}}$
Pearson correlation	$u_{Cor}(x,y) = 1 - \frac{\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{p}(x_j - \overline{x})^2}\sqrt{\sum_{j=1}^{p}(y_j - \overline{y})^2}} = 1 - \frac{<x-\overline{x}, y-\overline{y}>}{\|x-\overline{x}\|\|y-\overline{y}\|}$

TAB. 1 – *Quelques mesures de proximité.*

Où, p désigne la dimension de l'espace, $x = (x_j)_{j=1,\ldots,p}$ et $y = (y_j)_{j=1,\ldots,p}$ deux points de R^p, $(\alpha_j)_{j=1,\ldots,p} \geq 0$, $\sum^{-1}$ la matrice inverse de variances et covariances, $\overline{x}$ et $\overline{y}$ les moyennes, σ_j^2 la variance et $\gamma > 0$.

Cet article est organisé comme suit. Dans la section 2, après avoir rappelé les notions de structure, de graphe et d'équivalence topologique, nous présentons la façon dont a été construite la matrice d'adjacence dans un but de discrimination, le choix de la mesure du degré d'équivalence topologique entre deux mesures de proximité ainsi que le critère de sélection de la "meilleure" mesure discriminante. Un exemple illustratif est commenté en section 3. Une conclusion et quelques perspectives de cette approche sont données en section 4.

Le tableau 1 présente quelques mesures de proximité classiques utilisées pour des données continues, définies sur R^p.

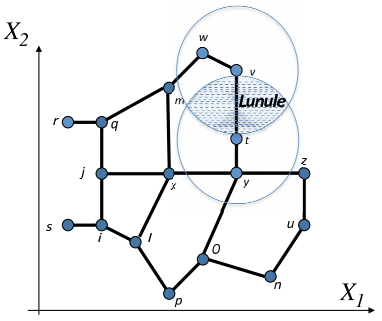
2 Equivalence topologique

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 correspondants 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 par la suite à comparer les mesures de proximité.

2.1 Graphe topologique

Sur un ensemble de points $x, y, z, \ldots$ de R^p, on peut, au moyen d'une mesure de proximité u 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'objets $E = \{x, y, z, \ldots\}$ de $n = |E|$ objets plongés dans R^p.

Ainsi, pour une mesure de proximité u donnée, nous pouvons construire un graphe de voisinage sur un ensemble d'individus dont les sommets sont les individus et les arêtes sont définis par une propriété de la relation de voisinage.



V_u	…	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 – *Graphe topologique GVR - Matrice d'adjacence binaire associée.*

De nombreuses définitions sont possibles pour construire cette relation binaire de voisinage. On peut, par exemple, choisir l'Arbre de Longueur Minimale (ALM) Kim et Lee (2003),

le Graphe de Gabriel (GG) Park et al. (2006), ou encore le Graphe des Voisins Relatifs (GVR) Toussaint (1980), dont tous les couples de points voisins vérifient la propriété suivante :

$$\begin{cases} V_u(x,y) = 1 & si \;\; u(x,y) \leq \max(u(x,z), u(y,z)) \; ; \; \forall z \in E - \{x,y\} \\ V_u(x,y) = 0 & sinon \end{cases} \quad (1)$$

c'est-à-dire, si les couples de points vérifient ou pas l'inégalité ultratriangulaire (1), condition ultramétrique.

La figure 1 montre, dans R^2 muni de la distance euclidienne $u(x,y) = u_E(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$, un exemple de graphe topologique GVR parfaitement défini par la matrice d'adjacence V_u associée, formée de 0 et de 1. Sur le plan géométrique, cela signifie que l'hyper-Lunule (intersection des deux hypersphères centrées sur les deux points) est vide.

2.2 Comparaison de mesures de proximité

On dispose de p variables quantitatives explicatives (prédicteurs) $\{x^j; j = 1,p\}$ et d'une variable qualitative cible à expliquer y, partition de $n = \sum_{k=1}^{q} n_k$ individus-objets en q modalités-groupes $\{G_k; k = 1,q\}$.

Pour toute mesure de proximité u_i donnée, on construit, selon la propriété (1), la matrice d'adjacence binaire globale V_{u_i} qui se présente comme une juxtaposition de q matrices d'adjacence symétriques Intra-groupe $\{V_{u_i}^k; k = 1,q\}$ et de $q(q-1)$ matrices d'adjacence Inter-groupes $\{V_{u_i}^{kl}; k \neq l; k,l = 1,q\}$:

$$\begin{cases} V_{u_i}^k(x,y) = 1 & si \; u_i(x,y) \leq \max(u_i(x,z), u_i(y,z)) \; ; \; \forall x,y,z \in G_k, \; z \neq x, \neq y \\ V_{u_i}^k(x,y) = 0 & sinon \end{cases}$$

$$\begin{cases} V_{u_i}^{kl}(x,y) = 1 & si \; u_i(x,y) \leq \max(u_i(x,z), u_i(y,z)) \; ; \; \forall x \in G_k, y \in G_l, \; z \in G_l, z \neq y \\ V_{u_i}^{kl}(x,y) = 0 & sinon \end{cases}$$

$$V_{u_i} = \begin{pmatrix} \mathrm{V}_{u_i}^1 & \cdots & \mathrm{V}_{u_i}^{lk} & \cdots & \mathrm{V}_{u_i}^{1q} \\ & \cdots & & & \\ \mathrm{V}_{u_i}^{k1} & \cdots & \mathrm{V}_{u_i}^{k} & \cdots & \mathrm{V}_{u_i}^{kq} \\ & & & \cdots & \\ \mathrm{V}_{u_i}^{q1} & \cdots & \mathrm{V}_{u_i}^{qk} & \cdots & \mathrm{V}_{u_i}^{q} \end{pmatrix}$$

A noter que la matrice d'adjacence partitionnée globale V_{u_i} ainsi construite, n'est pas symétrique. En effet, pour deux objets $x \in G_k$ et $y \in G_l$, les valeurs binaires d'adjacence $V_{u_i}^{kl}(x,y)$ et $V_{u_i}^{lk}(y,x)$ peuvent être différentes.

• Le premier objectif est de regrouper les différentes mesures de proximité considérées, selon leur similitude topologique pour mieux visualiser leur ressemblance dans un contexte de discrimination.

Pour mesurer le degré d'équivalence topologique de discrimination 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 pas. Le degré d'équivalence topologique entre deux mesures de proximité est mesuré par la quantité :

$$S(V_{u_i}, V_{u_j}) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n} \delta_{kl}}{n^2} \qquad \text{avec} \quad \delta_{kl} = \begin{cases} 1 \text{ si } \; V_{u_i}(k,l) = V_{u_j}(k,l) \\ 0 \text{ sinon.} \end{cases}$$

• Le second objectif consiste à établir un critère d'aide à la sélection de la "meilleure" mesure de proximité ; celle qui parmi toutes les mesures considérées, discrimine au mieux les q groupes.

On note, $V_{u^*} = diag(1_{G_1}, \ldots, 1_{G_k}, \ldots, 1_{G_q})$ la matrice d'adjacence symétrique bloc-diagonale de référence "discrimination parfaite" associée à la mesure de proximité inconnue, notée u^*, où, 1_{n_k} désigne le vecteur d'ordre n_k dont toutes les composantes sont égales à 1 et $1_{G_k} = 1_{n_k}\,{}^t1_{n_k}$, la matrice carrée d'ordre n_k dont tous les éléments sont égaux à 1.

$$V_{u^*} = \begin{pmatrix} 1_{G_1} & & & & \\ 0 & \cdots & & & \\ 0 & 0 & 1_{G_k} & & \\ 0 & 0 & 0 & \cdots & \\ 0 & 0 & 0 & 0 & 1_{G_q} \end{pmatrix}$$

On peut ainsi établir le degré d'équivalence topologique de discrimination $S(V_{u_i}, V_{u^*})$ entre chaque mesure de proximité u_i considérée et la mesure de référence u^*.

Enfin, afin d'évaluer autrement le choix de la "meilleure" mesure de proximité discriminante proposée par cette approche, nous avons appliqué *a posteriori* une technique de classement par SVM Multiclasses (MSVM) sur la matrice d'adjacence associée à chacune des mesures de proximité considérée y compris à la mesure de référence u^*.

3 Exemple d'application

Pour illustrer notre approche, nous considérons ici un jeu de données bien connu et relativement simple, celui des Iris Fisher (1936); Anderson (1935). Ces données ont été proposées comme données de référence pour l'analyse discriminante et la classification par le statisticien Ronald Aylmer Fisher en 1933. Les données complètes se trouvent notamment dans UCI (2013). Quatre variables (longueur et largeur des sépales et pétales) ont été observées sur 50 fleurs de chacune des 3 espèces d'Iris (Iris Setosa, Iris Virginica, Iris Versicolor).

3.1 Comparaison et classement des mesures de proximité

Les principaux résultats de l'approche proposée sont présentés dans les tableaux et le graphique suivants. Ils permettent de visualiser les mesures qui sont proches les unes des autres selon l'objectif de discrimination.

Pour ce jeu de données, le tableau 2 récapitule les similarités entre les 8 mesures de proximité et montre que la mesure u_{Tch} de Tchebychev est la plus proche de la mesure u^* de référence.

Une Analyse en Composantes Principales (ACP) suivie d'une Classification Hiérarchique Ascendante (CHA) ont été effectuées à partir de la matrice de similarités entre les 8 mesures de proximité considérées, afin de les partitionner dans des groupes homogènes et de visualiser leurs ressemblances.

L'application d'un algorithme de construction d'une CHA selon le critère de Ward, Ward Jr (1963), permet d'obtenir le dendogramme de la figure 2. Le vecteur similarités $S(V_{u_i}, V_{u^*})$ de

S	u_E	u_{Mah}	u_{Man}	u_{Tch}	u_{Cos}	u_{NE}	$u_{Min_{\gamma=5}}$	u_{Cor}
u_E	1							
u_{Mah}	0.953	1						
u_{Man}	0.977	0.947	1					
u_{Tch}	0.968	0.934	0.949	1				
u_{Cos}	0.955	0.946	0.949	0.939	1			
u_{NE}	0.968	0.956	0.969	0.945	0.950	1		
$u_{Min_{\gamma=5}}$	0.992	0.951	0.971	0.975	0.953	0.965	1	
u_{Cor}	0.949	0.943	0.944	0.930	0.966	0.946	0.948	1
u^*	0.675	0.673	0.678	**0.681**	0.675	0.674	0.675	0.673

TAB. 2 – *Equivalences topologiques - Similarités $S(V_{u_i}, V_{u_j})$ et $S(V_{u_j}, V_{u^*})$.*

la mesure de référence u^* avec les mesures de proximité considérées est positionné en élément illustratif dans l'analyse.

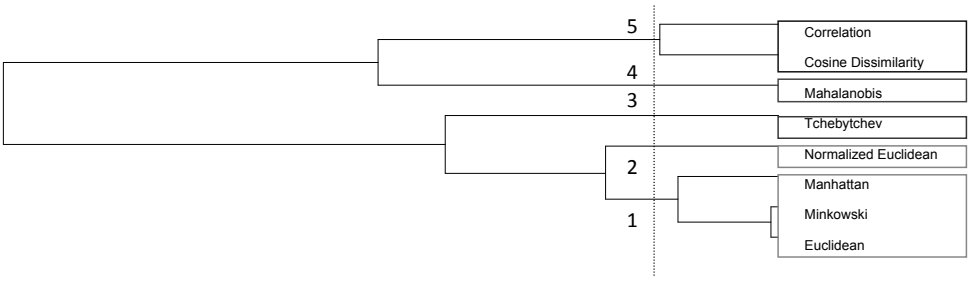


FIG. 2 – *Dendogramme - Structure topologique GVR.*

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Effectif	3	1	1	1	2
Mesures actives	$u_E,\ u_{Min},\ u_{Man}$	u_{NE}	u_{Tch}	u_{Mah}	u_{Cos}, u_{Cor}
Mesure illustrative			$\mathbf{u}^*$		

TAB. 3 – *Classement de la mesure de référence.*

Au vu des résultats présentés dans le tableau 3 de la partition en 5 classes de mesures de proximité, la mesure de référence inconnue u^*, projetée en élément supplémentaire, serait donc plus proche des mesures de la classe 3, c'est-à-dire, de la mesure de Tchebychev u_{Tch} qui serait pour ces données, la "meilleure" mesure de proximité parmi les 8 mesures considérées. Ce résultat confirme celui constaté dans le tableau 2, à savoir, une plus grande similarité $S(V_{u_{Tch}}, V_{u^*}) = 68.10\%$ de la mesure de Tchebychev avec celle de référence u^*.

3.2 Les mesures discriminantes selon les MSVM

Cette partie consiste à valider les résultats du choix de la meilleure mesure au vu de la matrice de référence *a posteriori* en utilisant les MSVM. Nous utilisons ici le modèle $MSVM_{LLW}$, Lee et al. (2004), considéré comme le plus fondé théoriquement du fait que sa solution donne un classifieur qui converge vers celui de Bayes.

Mesure	Erreur d'apprentissage (%)	Matrice de confusion
u_E	**0**	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 50 \end{pmatrix}$
u_{Mah}	0.66	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 0 & 50 \end{pmatrix}$
u_{Man}	0.66	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 0 & 50 \end{pmatrix}$
u_{Tch}	**0**	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 50 \end{pmatrix}$
u_{Cos}	0.66	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 0 & 50 \end{pmatrix}$
u_{NE}	1.33	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 2 & 48 \end{pmatrix}$
$u_{Min_\gamma=5}$	1.33	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 1 & 49 \end{pmatrix}$
u_{Cor}	1.33	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 1 & 49 \end{pmatrix}$
u^*	**0**	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 50 \end{pmatrix}$

TAB. 4 – *Principaux résultats du modèle MSVM*

Travailler avec le modèle $MSVM_{LLW}$ implique le choix des valeurs optimales de ses paramètres : C, représentant le poids des erreurs d'apprentissage, et le ou les paramètres de la fonction noyau si nous décidons de changer l'espace de données.

Dans le cas des données des Iris de Fisher, nous avons choisi de travailler dans l'espace d'origine des données et donc d'utiliser un noyau linéaire. Le seul paramètre à optimiser est C.

Pour ce faire, nous allons appliquer une des techniques de sélection du modèle consistant à tester plusieurs valeurs du paramètre et à choisir celle qui minimise l'erreur test calculée par

validation croisée. Dans cet exemple, nous testons 10 valeurs du paramètres (entre 1 et 100) pour toutes les bases de données. Aprés simulations, la valeur choisie est $C = 1$.

Les principaux résultats du modèle $MSVM_{LLW}$, appliqué sur chacune des matrices d'adjacence induites par les mesures de proximité considérées, sont présentés dans le tableau 4.

Le meilleur taux d'erreur est celui donné par les mesures de Tchebechev u_{Tch} et Euclidienne u_E, qui est aussi celui enregistré pour la matrice de référence.

L'application du modèle MSVM montre que les mesures de proximité de Tchebechev u_{Tch} et Euclidienne u_E sont les plus adaptées pour différencier et séparer au mieux les 3 espèces d'Iris. Ce résultat confirme celui obtenu précédemment, à savoir le choix de la mesure de Tchebychev u_{Tch} comme la mesure plus proche, parmi les huit mesures considérées, de la mesure de référence et donc la plus discriminante.

3.3 Expérimentations

Nous avons procédé à des expérimentations sur d'autres jeux de données afin d'essayer d'évaluer l'effet des données, de leur taille et/ou de leur dimension sur les résultats de la classification des mesures de proximité toujours dans un but de discrimination. Est-ce que, par exemple, les mesures de proximité se regroupent différemment selon le jeu de données utilisé ? Selon la taille de l'échantillon et/ou le nombre de variables explicatives considérées dans un même ensemble de données ?

Pour répondre à ces questions, nous avons donc appliqué l'approche proposée sur différents jeux de données, présentés dans le tableau 5, qui proviennent tous du référentiel UCI (2013). L'objectif est de comparer les résultats des classifications des mesures de proximité de toutes ces expérimentations ainsi que la "meilleure" mesure discriminante proposée pour chacun de ces jeux de données.

Etant donné un ensemble de données explicatives $X_{(n,p)}$ à n objets et p variables, et une variable à expliquer ou à discriminer Y_q à q modalités-classes.

Pour analyser l'effet du changement de dimension, nous avons considéré le jeu de données "Waveform Database Generator" pour générer 3 échantillons $n^{\circ}4$ de taille $n = 2000$ objets et de dimension p égale respectivement à 40, 20 et à 10 variables.

De même, pour évaluer l'effet du changement de la taille de l'échantillon, nous avons également généré 3 autres échantillons $n^{\circ}5$ de taille n égale respectivement à 3000, 1500 et à 500 objets et de même dimension p égale à 30 variables.

Les principaux résultats de ces expérimentations, à savoir les équivalences topologiques des mesures de proximité discriminantes et l'affectation de la mesure de référence u^* dans la classe la plus proche, sont présentés dans le tableau 6.

Pour chacune de ces expérimentations, nous avons retenu une partition en cinq classes de mesures de proximité afin de les comparer et de bien distinguer les mesures de la classe d'appartenance de la mesure de référence, c'est-à-dire les mesures les plus discriminantes.

Les regroupements des mesures de proximité obtenus pour les trois jeux de données $n^{\circ}4$ sont pratiquement identiques, il n'y a donc pas vraiment d'effet de la dimension.

n°	Nom	$X_{(n\times p)}$	$Y_{(q)}$
1	Iris	150 × 4	3
2	Wine	178 × 13	3
3	Wine Quality	3000 × 11	2
4_1	Waveform Database Generator	2000 × 40	3
4_2	Waveform Database Generator	2000 × 20	3
4_3	Waveform Database Generator	2000 × 10	3
5_1	Waveform Database Generator	3000 × 30	3
5_2	Waveform Database Generator	1500 × 30	3
5_3	Waveform Database Generator	500 × 30	3

TAB. 5 – *Jeux de données.*

n°	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
1	u_{Cos}, u_{Cor}	$u_E,\ u_{Min},\ u_{Man}$	u_{Mah}	u_{NE}	$u_{Tch},\ \mathbf{u}^*$
2	u_{Cos}, u_{Cor}	$u_E,\ u_{Min},\ u_{Tch}$	u_{Mah} , $\mathbf{u}^*$	u_{NE}	u_{Man}
3	u_{Cos}, u_{Cor}	$u_E,\ u_{Min},\ u_{Man}$	u_{Mah}	$u_{NE}, \mathbf{u}^*$	u_{Tch}
4_1	u_{Cos}, u_{Cor}, u_E	$u_{Man},\ u_{NE}$	u_{Mah}	u_{Min}	$u_{Tch},\ \mathbf{u}^*$
4_2	$u_{Cos}, u_{Cor},\ u_E,\ u_{NE}$	u_{Man}	u_{Mah}	u_{Min}	$u_{Tch},\ \mathbf{u}^*$
4_3	u_{Cos}, u_{Cor}	$u_E,\ u_{Man}, u_{NE}$	u_{Mah}	u_{Min}	$u_{Tch},\ \mathbf{u}^*$
5_1	u_{Cos}, u_{Cor}, u_E	$u_{Man},\ u_{NE}$	u_{Mah}	u_{Min}	$u_{Tch},\ \mathbf{u}^*$
5_2	u_{Cos}, u_{Cor}, u_E	$u_{Man},\ u_{NE}$	u_{Mah}	u_{Min}	$u_{Tch},\ \mathbf{u}^*$
5_3	u_{Cos}, u_{Cor}, u_E	$u_{Man},\ u_{NE}$	u_{Mah}	u_{Min}	$u_{Tch},\ \mathbf{u}^*$

TAB. 6 – *Regroupements et affectation de la mesure de référence* u^*.

Quant aux regroupements des mesures de proximité des trois jeux de données $n^\circ 5$, ils sont quasiment identiques, il n'y a donc pas du tout d'effet de la taille des échantillons.

A noter que pour tous les échantillons $n^\circ 4$ et $n^\circ 5$, tous générés du même jeu de données "Waveform Database Generator", la mesure de référence idéale u^* pour la discrimination est proche de la même mesure de proximité à savoir ici la mesure u_{Tch} de Tchebychev. Ce résultat montre qu'il n'y a ni effet de taille ni effet de dimensionalité sur le résultat du choix de la meilleure mesure discriminante.

Relativement à l'ensemble des expérimentations, on peut constater un léger changement dans les regroupements des mesures de proximité. Cependant, on peut noter aussi des équivalences entre certaines mesures telles que $\{u_{Cos}, u_{Cor}, u_E\}$ et $\{u_{NE}, u_{Man}\}$. D'autres encore restent isolées telles que $\{u_{Tch}\}$, $\{u_{Mah}\}$ ou encore $\{u_{Min}\}$.

4 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 dans un contexte de discrimination.

Cette approche topologique est basée sur la notion de graphe de voisinage induit par la mesure 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 données qualitatives 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, évaluer le degré d'équivalence topologique de discrimination entre deux mesures de proximité en appliquant le test non paramètrique du coefficient de concordance de Kappa, calculé à partir des matrices d'adjacence associées, Abdesselam et Zighed (2011). Cela va permettre de donner une signification statistique du degré de concordance entre les deux matrices de ressemblance et de valider ou pas l'équivalence topologique de discrimination, c'est-à-dire si vraiment elles induisent ou pas la même structure de voisinage sur les groupes d'objets à séparer.

Enfin, les expérimentations menées sur différents jeux de données ont montré qu'il n'y pas du tout d'effet de la dimension et pas vraiment d'effet de la taille de l'échantillon aussi bien sur les regroupement des mesures de proximité que sur le résultat du choix de la meilleure mesure discriminante.

Références

Abdesselam, R. (2014). Proximity measures in topological structure for discrimination. *In a Book Series SMTDA-2014, 3nd Stochastic Modeling Techniques and Data Analysis, International Conference, Lisbon, Portugal, C.H. Skiadas (Ed), ISAST*, 599–606.

Abdesselam, R. et D. Zighed (2011). Comparaison topologique de mesures de proximité. *Actes des XVIIIème Rencontres de la Société Francophone de Classification*, 79–82.

Anderson, E. (1935). The irises of the gaspe peninsula. *Bulletin of the American Iris Society, 59*, 2–5.

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.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, Part II, 7*, 179–188.

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

Lee, Y., Y. Lin, et G. Wahba (2004). Multicategory support vector machines,theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association, 465*, 67–81.

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

Liu, H., D. Song, S. Ruger, R. Hu, et V. Uren (2008). 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.

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.

Richter, M. (1992). Classification and learning of similarity measures. *Proceedings der Jahrestagung der Gesellschaft fur 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.

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

UCI (2013). UCI machine learning repository, [http://archive.ics.uci.edu/ml]. irvine, CA: University of california, school of information and computer science.

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.

Zighed, D., R. Abdesselam, et A. Hadgu (2012). Topological comparisons of proximity measures. *The 16th PAKDD 2012 Conference. In P.-N. Tan et al. (Eds.) Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg*, 379–391.

Summary

The results of any operation of clustering or classifying of objects are strongly depend on the proximity measure chosen. The user has to select one measure among many existing proximity measures. Yet according to the notion of topological equivalence chosen, some are more or less equivalent. In this paper, we propose a new approach to comparing and classifying

proximity measures in a topological structure and a goal of discrimination. The concept of topological equivalence uses the structure of local neighborhood.

Then we propose to define the topological equivalence between two proximity measures, in the context of discrimination, through the topological structure induced by each measure. We also propose a criterion for choosing the "best" measure adapted to data considered among some of the most used proximity measures for quantitative data. The choice of the "best" discriminating proximity measure can be verified retrospectively by a supervised learning method type SVM, discriminant analysis or Logistic regression applied in a topological context.

The principle of the proposed approach is illustrated using a real quantitative data example with eight conventional proximity measures of literature. Experiments have evaluated the performance of this discriminant topological approach in terms of size and/or dimension of the relevant data and of selecting the "best" discriminant proximity measure.