Degrés d'équivalence de mesures de comparaison pour données binaires et pour données numériques

Marie-Jeanne Lesot*, Maria Rifqi*

*LIP6 - Université Pierre et Marie Curie-Paris6, UMR7606
4, place Jussieu 75252 Paris cedex 05
prénom.nom@lip6.fr

Résumé. Afin d'aider au choix d'une mesure pour comparer des données, problème au cœur de la conception de systèmes dans les domaines de la fouille de données, l'apprentissage automatique ou la recherche d'information, nous comparons les mesures les plus courantes selon l'ordre qu'elles induisent sur les données et nous quantifions leur accord par des degrés d'équivalence. Nous proposons une étude systématique des mesures de comparaison appliquées aux données binaires et aux données numériques, en examinant les principales mesures de similarité, distance et produits scalaires. Nous établissons leurs degrés d'équivalence, en considérant des bases de données artificielles et réelles et identifions des mesures équivalentes et quasi-équivalentes, qui peuvent être considérées comme redondantes dans un cadre de recherche d'information.

1 Introduction

Les mesures de comparaison, qui regroupent similarités et distances, sont des fonctions qui quantifient la ressemblance ou, de façon duale, la dissimilarité entre objets : elles prennent en argument des paires d'objets et renvoient des valeurs numériques qui sont d'autant plus élevées, dans le cas des mesures de similarité, d'autant plus faibles dans le cas des mesures de distance, que les objets sont proches. Il existe de très nombreuses mesures de comparaison, qui diffèrent selon le type de données auxquelles elles s'appliquent (données binaires, numériques ou structurées par exemple) ainsi que selon leurs sémantiques (Lesot et al., 2009).

Dans cet article, nous nous intéressons aux cas des données binaires et des données numériques. Les premières sont décrites dans l'univers $\{0,1\}^q$ et sont également appelées données présence/absence ou données ensemblistes : la valeur d'un attribut indique si la propriété correspondante est présente ou non dans la donnée considérée ; la donnée peut aussi être décrite par l'ensemble des propriétés qu'elle présente. Les données numériques sont des données vectorielles décrites dans $\mathbb{R}^q$.

Le choix d'une mesure de comparaison est au cœur de la conception de systèmes dans les domaines de la fouille de données, l'apprentissage automatique ou la recherche d'information. Suivant les tâches considérées, différentes contraintes sur la mesure doivent être prises en compte. On peut distinguer deux cadres de sélection des mesures de comparaison, selon qu'on s'intéresse aux valeurs numériques qu'elles fournissent ou aux ordres qu'elles induisent. Dans

FIG. 1 – *Listes d'images classées par similarité décroissante avec l'image requête, pour quatre mesures de comparaison. Les deux premières listes sont strictement identiques, de même que les deux dernières. Dans leur ensemble, elles diffèrent par la quatrième image renvoyée.*

le premier cas, la comparaison des mesures peut être basée sur le pouvoir de discrimination, qui évalue la sensibilité locale des mesures aux variations des objets comparés (Rifqi et al., 2000, 2003), ou sur le degré de sévérité, qui quantifie la ressemblance minimale nécessaire à l'obtention d'un niveau de similarité donné (Lesot et al., 2009).

Nous nous plaçons, dans cet article, dans le cas où ce sont les ordres induits par les mesures qui importent et non les valeurs numériques qu'elles prennent. Cette contrainte est fréquente et illustrée par exemple dans le cadre de la recherche d'information : les résultats fournis par les moteurs de recherche prennent la forme de listes de documents ordonnés par pertinence décroissante, la pertinence étant le plus souvent calculée comme la similarité entre le document candidat et la requête. Ainsi, seul l'ordre induit est pris en compte, non les valeurs de similarité : les critères d'évaluation classiques, basés sur le rappel et la précision, ne dépendent que de l'ordre des résultats. Aussi, il n'est pas utile d'hésiter entre des mesures qui donnent le même classement des données.

Ce principe est illustré sur la figure 1, qui montre les 4 premières images retournées pour

la requête indiquée dans la partie supérieure, dans un corpus constitué des images issues des pages tourisme du site Wikipedia français. Quatre mesures distinctes (voir le rappel des formules dans la section 3.3.2) sont utilisées, chacune correspondant à une colonne de la partie résultats : on observe que la distance euclidienne et le noyau gaussien (2 premières colonnes) renvoient la même liste d'images, de même que le produit scalaire euclidien et le noyau polynomial (2 dernières colonnes). Ainsi, même si ces quatre mesures donnent des scores de ressemblance différents entre la requête et les images de la base, seules deux listes différentes sont obtenues : la question du choix de la mesure se pose seulement entre 2 catégories de mesures, et non 4 mesures. De plus, entre ces deux groupes, une seule différence, concernant la quatrième image se produit. Ainsi, on peut considérer qu'elles sont globalement redondantes, et que la question du choix de la mesure ne se pose pas réellement.

De façon analogue, dans le cadre de l'apprentissage automatique, les algorithmes de clustering hiérarchique avec chaînage simple ou avec chaînage complet, de même que l'algorithme de clustering monotone invariant (Janowitz, 1979) exploitent l'ordre induit par les mesures de comparaison des données, et non leurs valeurs numériques.

Aussi, dans ce cadre, le choix des mesures de comparaison peut être basé sur les notions d'équivalence et de classes d'équivalence : ces dernières, introduites initialement pour les mesures de similarité pour données ensemblistes (Lerman, 1967; Baulieu, 1989; Batagelj et Bren, 1995; Omhover et al., 2006), établissent des catégories de mesures selon l'ordre qu'elles induisent, de telle sorte que l'appartenance à une même classe garantit des résultats ordonnés identiques. La notion d'équivalence a été raffinée par la définition de degrés d'équivalence permettant d'examiner plus finement les écarts entre mesures non équivalentes et de définir des mesures quasi-équivalentes : ces degrés quantifient le désaccord entre les ordres induits et indiquent de la sorte des proximités plus riches entre mesures (Rifqi et al., 2008). Ils renseignent sur les redondances et permettent de guider le choix de mesures en fonction des attentes de variabilité souhaitée ou tolérée dans les résultats fournis par les mesures.

Dans cet article, nous proposons une étude systématique de ces propriétés d'équivalence et de quasi-équivalence pour les mesures de comparaison appliquées à deux types de données, les données binaires et les données numériques : nous examinons les principales mesures de similarité, distance et produits scalaires et établissons leurs degrés d'équivalence que nous calculons pour des données artificielles et réelles. Les deux corpus permettent en particulier d'étudier les effets, sur les résultats d'équivalence, d'éventuelles configurations spécifiques des données, par exemple des densités variables ou des structures en clusters.

L'article est organisé de la façon suivante : la section 2 rappelle les définitions d'équivalence et de degrés d'équivalence entre mesures de comparaison, la section 3 détaille le protocole expérimental et les bases de données, artificielles et réelles, utilisées pour l'étude proposée. Les sections 4 et 5 analysent les résultats obtenus pour les données binaires et numériques respectivement, à la fois d'un point de vue théorique et d'un point de vue expérimental.

2 Degrés d'équivalence entre mesures de comparaison

2.1 Propriétés générales des mesures de comparaison

De façon générale, en notant $\mathcal{X}$ un espace de données, une mesure de comparaison est une fonction $m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ qui associe à une paire d'objets une valeur numérique quantifiant

leur ressemblance ou, de façon duale, leur dissimilarité.

Dans le cas de données ensemblistes, ou données présence/absence, il existe une théorisation des mesures de similarité, notamment en termes de propriétés attendues (Bouchon-Meunier et al., 1996) : une mesure de similarité est habituellement définie comme une fonction $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ possédant les propriétés de

- positivité : $\forall x, y \in \mathcal{X}$, $s(x, y) \geq 0$
- symétrie : $\forall x, y \in \mathcal{X}$, $s(x, y) = s(y, x)$
- maximalité : $\forall x, y \in \mathcal{X}$, $s(x, x) \geq s(x, y)$

Toutefois, Tversky (1977) rejette par exemple la propriété de symétrie, soulignant que dans des comparaisons du type "x ressemble à y", les deux éléments comparés ne jouent pas le même rôle : il faut distinguer y, objet de référence, auquel x est comparé. La valeur de ressemblance peut donc être différente de celle associée à l'assertion "y ressemble à x". Par ailleurs, d'autres propriétés sont parfois exigées, comme la normalisation des valeurs obtenues. Cette dernière doit typiquement garantir des valeurs dans l'intervalle [0, 1].

En ce qui concerne les données numériques, les mesures de comparaison sont de deux types, distances et produits scalaires. Les distances ont une sémantique opposée à celle des similarités, et un sens de variation opposé en fonction de la ressemblance. On peut les transformer en mesures de similarité si on applique une fonction décroissante (cf section 3.3.2). Dans ce qui suit, quand des comparaisons faisant intervenir des distances sont mentionnées, il est entendu qu'elles se basent sur l'ordre induit par $-d$.

Les distances possèdent des propriétés équivalentes à celles des mesures pour données ensemblistes précédentes, à l'inversion de sens près pour la propriété de maximalité :

- positivité : $\forall x, y \in \mathcal{X}$, $d(x, y) \geq 0$
- symétrie : $\forall x, y \in \mathcal{X}$, $d(x, y) = d(y, x)$
- minimalité : $\forall x, y \in \mathcal{X}$, $d(x, y) = 0 \Leftrightarrow x = y$, qui implique $d(x, x) \leq d(x, y)$

Elles vérifient de plus l'inégalité triangulaire, $d(x, y) \leq d(x, z) + d(z, y)$, alors que les similarités min-transitives, telles que $s(x, z) \geq \min(s(x, y), s(y, z))$, forment une catégorie spécifique au sein des mesures ensemblistes.

Les produits scalaires sont plus proches des mesures de similarité, car ils sont également des fonctions croissantes de la ressemblance, bien qu'ils utilisent une définition différente de ressemblance. Il faut toutefois souligner qu'ils présentent la différence essentielle de ne pas posséder la propriété de maximalité : on a par exemple $\langle x, 2x \rangle > \langle x, x \rangle$.

Nous présentons ici des outils de comparaison de ces mesures basés sur l'ordre qu'elles induisent, en décrivant successivement les notions d'équivalence et de quasi-équivalence, cette dernière étant quantifiée par les degrés d'équivalence.

2.2 Définition d'équivalence d'ordres induits

La comparaison théorique entre mesures de comparaison a été étudiée par de nombreux auteurs (Lerman, 1967; Baulieu, 1989; Batagelj et Bren, 1995; Omhover et al., 2006), initialement dans le cas des mesures de similarité pour données ensemblistes. Elle a conduit à la

notion d'équivalence, définie de la façon suivante :

$$\text{deux mesures } m_1 \text{ et } m_2 \text{ sont } \textit{équivalentes} \text{ si et seulement si}$$
$$\forall x,y,z,t \in \mathcal{X} \quad \begin{cases} m_1(x,y) < m_1(z,t) \Longleftrightarrow m_2(x,y) < m_2(z,t) \\ m_1(x,y) = m_1(z,t) \Longleftrightarrow m_2(x,y) = m_2(z,t) \end{cases}$$

Cette propriété doit être vérifiée indépendamment des données considérées, c'est-à-dire quel que soit $\mathcal{X}$.

Cette définition implique que deux mesures équivalentes induisent le même ordre : ainsi, si on calcule les valeurs de similarité entre une donnée de référence et un ensemble de données, c'est-à-dire si on considère $x = y = x^*$, z et t restant variables dans $\mathcal{X}$, la définition précédente impose la préservation de l'ordre relatif entre les valeurs $m_1(x^*, z)$ et $m_1(x^*, t)$. Le classement des données par similarité décroissante à la requête x^* doit donc être le même pour les deux mesures. Cette notion est donc centrale en particulier pour les applications telles que les moteurs de recherche.

La notion d'équivalence peut également être définie directement en fonction des mesures elles-mêmes, sans passer explicitement par la comparaison de données : la définition précédente est équivalente à la définition ci-dessous (Batagelj et Bren, 1995; Omhover et al., 2006).

$$\text{deux mesures } m_1 \text{ et } m_2 \text{ sont } \textit{équivalentes} \text{ si et seulement si}$$
$$\exists f : Im(m_1) \to Im(m_2) \text{ strictement croissante telle que } m_2 = f \circ m_1$$

où $Im(m) \subset \mathbb{R}$ est l'ensemble des valeurs que peut prendre la mesure m.

2.3 Définition de degrés d'équivalence d'ordres induits

Afin d'établir des distinctions entre les mesures qui ne vérifient pas la propriété d'équivalence précédente, et d'examiner plus en détails ces mesures non équivalentes, il a été proposé de définir des *degrés d'équivalence* pour quantifier le désaccord entre mesures (Rifqi et al., 2008) : ces désaccords proviennent de quadruplets de données tels que $m_1(x,y) < m_1(z,t)$ mais $m_2(x,y) > m_2(z,t)$, correspondant à des inversions, ou tels que l'on ait égalité pour l'une des deux mesures, et non pour l'autre, par exemple $m_1(x,y) = m_1(z,t)$ mais $m_2(x,y) \neq m_2(z,t)$.

Le nombre de telles inversions est un critère d'importance : deux mesures produisant des ordres opposés l'un à l'autre, c'est-à-dire tels que tous les quadruplets sont inversés, sont "plus fortement" non équivalentes que deux mesures qui ne conduisent qu'à quelques inversions. Il est de plus intéressant de tenir compte des positions de ces inversions : si elles se produisent pour des valeurs de similarité faibles, les mesures peuvent être considérées comme plus équivalentes que si elles sont en désaccord pour les valeurs élevées. Pour les moteurs de recherche par exemple, le plus souvent, seuls les premiers résultats sont pris en compte, et des inversions pour les derniers résultats ne seront pas même remarquées.

Aussi, il a été proposé de définir des degrés d'équivalence, qui tiennent compte à la fois du nombre d'inversions et de leurs positions (Rifqi et al., 2008), en utilisant le coefficient de Kendall généralisé (Fagin et al., 2003, 2004) appliqué à des listes ordonnées tronquées à leurs premiers éléments.

Coefficient de Kendall généralisé Le coefficient de Kendall permet de mesurer la corrélation entre deux ordres, en calculant la proportion d'inversions qu'ils présentent. Les généralisations proposées par Fagin et al. (2003, 2004), détaillées ci-dessous, permettent de tenir compte des ex-aequo et de traiter des classements ne comportant pas les mêmes éléments. Ce dernier cas peut se produire quand on tronque les listes ordonnées, pour ne considérer que leurs premiers éléments.

Plus formellement, notons r_1 et r_2 deux ordres à comparer, définis sur un ensemble $\mathcal{E}$ de n éléments : $r(i)$ indique le rang du i-ème objet selon l'ordre r.

Le coefficient de Kendall généralisé associe à chaque paire d'éléments $(i, j) \in \mathcal{E}^2$ une pénalité $P_{r_1,r_2}(i, j)$, puis est calculé comme la somme des pénalités rapportée au nombre de comparaisons effectuées, c'est-à-dire

$$K_{p,p'}(r_1, r_2) = \frac{2}{n(n-1)} \sum_{i \neq j} P_{r_1,r_2}(i, j) \tag{1}$$

Le calcul des pénalités dépend de deux paramètres p et p' et distingue quatre valeurs selon les propriétés de la paire (i, j)

- $P_{r_1,r_2}(i, j) = 0$ s'il s'agit d'une paire dite concordante, c'est-à-dire si l'on a $r_1(i) < r_1(j)$ et $r_2(i) < r_2(j)$ ou si $r_1(i) > r_1(j)$ et $r_2(i) > r_2(j)$ ou si $r_1(i) = r_1(j)$ et $r_2(i) = r_2(j)$.
- $P_{r_1,r_2}(i, j) = 1$ si la paire est discordante, c'est-à-dire si les classements donnés par r_1 et r_2 sont inversés.
- $P_{r_1,r_2}(i, j) = p \in [0, 1]$ si la paire est ex-aequo dans l'un des classements, mais non dans l'autre, c'est-à-dire si $r_1(i) = r_1(j)$ mais $r_2(i) \neq r_2(j)$, ou vice-versa.
- le dernier cas enfin concerne les paires manquantes, c'est-à-dire les cas où la paire (i, j) est présente dans l'une des listes, mais non dans l'autre. Ce cas peut se produire lorsqu'on considère des listes tronquées, réduites à leurs premiers éléments. Comme illustré sur la figure 2, la pénalité vaut alors $P_{r_1,r_2}(i, j) = p' \in [0, 1]$ si i et j sont tous deux absents de la seconde liste ; si un seul d'entre eux est absent, la paire est traitée comme une paire normale et la pénalité vaut 0, 1 ou p selon qu'elle est concordante, discordante ou ex-aequo.

Le coefficient de Kendall généralisé vaut alors 0 si les deux ordres comparés sont identiques, et 1 si les deux ordres sont inversés.

Degrés d'équivalence Le degré d'équivalence entre deux mesures de comparaison m_1 et m_2 est alors défini dans un cadre de recherche d'information et de comparaison avec une donnée requête, comme le coefficient de Kendall généralisé appliqué aux ordres induits par les mesures (Rifqi et al., 2008). Plus précisément, on considère que les données sont comparées à une donnée de référence x^*, et $r(i)$ désigne le rang de la i-ème donnée dans cette comparaison, c'est-à-dire dans le classement selon $s(x^*, x_i)$. On tronque ensuite les listes obtenues pour ne considérer que les objets de rang inférieur à un paramètre k, et on note r^k l'ordre restreint. Le degré d'équivalence est alors défini comme

$$d_k(m_1, m_2) = 1 - K_{0.5,1}(r_1^k, r_2^k) \tag{2}$$

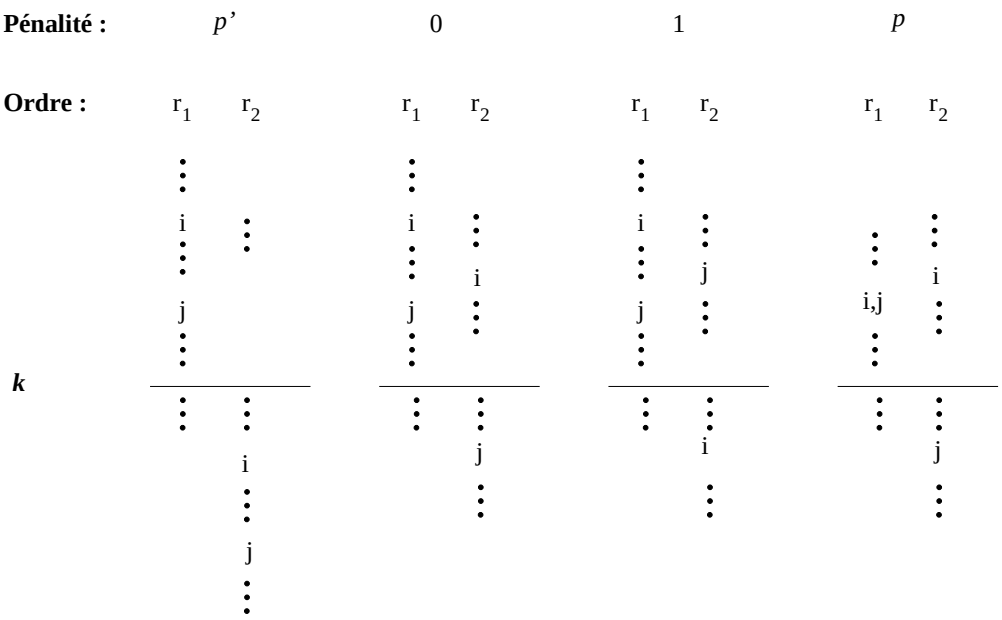

FIG. 2 – *Pénalité $P_{r_1,r_2}(i,j)$ dans le cas de paires manquantes, c'est-à-dire pour les différentes configurations où la paire (i,j) est absente de l'ordre r_2 tronqué au rang k.*

La transformation $1-K$ permet de définir un degré qui est maximal (valant 1) pour les mesures équivalentes, et minimal (valant 0) pour des mesures conduisant à des classements opposés.

Les paramètres p et p' influencent les résultats obtenus en déterminant la tolérance à l'égalité et à l'absence. Nous faisons les choix suivants : $p = 0,5$ et $p' = 1$. En effet, pour les paires ex-aequo, nous considérons qu'en choisissant arbitrairement un ordre strict pour les départager, on a une chance sur deux de produire le même ordre que dans l'autre liste, conduisant à une chance sur deux d'avoir deux classements différents. Enfin nous considérons que les paires manquantes indiquent une différence majeure entre les deux classements, et peuvent être pénalisées autant qu'une paire discordante, c'est-à-dire $p' = 1$.

3 Cadre expérimental

3.1 Protocole

Pour calculer les degrés d'équivalence entre deux mesures m_1 et m_2, nous appliquons le protocole expérimental suivant, adapté du principe des moteurs de recherche : étant donné une requête x^*, toutes les données d'une base sont classées par ordre de similarité décroissante, en utilisant les mesures de comparaison étudiées. Les classements obtenus sont alors comparés, soit intégralement, soit en les restreignant aux objets de rang inférieur ou égal à k, paramètre fixé. Les degrés d'équivalence finals sont les moyennes des valeurs obtenues lorsqu'on fait varier la donnée requête x^* sur toutes les données de la base considérée.

Les expériences sont réalisées avec des données réelles et des données artificielles pour chacun des deux types de données, binaires (appartenant à $\mathcal{X} = \{0,1\}^q$) et numériques (appartenant à $\mathcal{X} = \mathbb{R}^q$). Comme détaillé ci-dessous dans la description des corpus pour chaque type, les données artificielles sont générées uniformément afin de pouvoir étudier, par comparaison avec les données réelles, les effets sur les résultats d'équivalence d'éventuelles configurations spécifiques des données, par exemple des densités variables ou des structures en clusters.

Outre les corpus de données, les sous-sections suivantes précisent également les mesures de comparaison étudiées, dans chacun des cas. Pour chaque type de données, nous incluons de plus, à titre de référence, une mesure qui génère des valeurs de similarité aléatoires pour tout couple de données.

3.2 Cas des données binaires

3.2.1 Corpus de test

Les données réelles binaires sont construites à partir du corpus du Challenge ImageCLEF (2008) qui contient 1827 images annotées selon leur contenu. Chaque image possède une ou plusieurs étiquettes, dont les valeurs possibles sont structurées dans une hiérarchie de concepts. Celle-ci comporte des objets (*buildings*, *vegetation*) ainsi que des propriétés (*outdoor*, *night* par exemple). Les annotations des images constituent donc des données binaires indiquant la présence ou l'absence de chaque concept dans l'image.

Nous avons supprimé les étiquettes en relation XOR (telles que *night*, qui exclut *day*, ou *outdoor*, qui exclut *indoor*), ainsi que des sous-catégories, qui impliquent nécessairement leurs sur-catégories, comme *sunny*, *partly cloudy* et *overcast* qui imposent *sky*, et *tree* qui implique *vegetation*. La base finale de données binaires comporte $q = 11$ attributs.

Les données artificielles sont générées aléatoirement dans le même espace selon une distribution uniforme, afin de tester les effets potentiels de configurations spécifiques des données. Elles sont constituées d'une grille régulière de $\{0,1\}^{11}$, qui conduit à $2^{11} = 2048$ points.

3.2.2 Mesures considérées

Etant donné deux objets x et y appartenant à $\{0,1\}^q$, on note $X = \{i|x_i = 1\}$ et $Y = \{i|y_i = 1\}$ les attributs présents dans chacun d'eux respectivement. Les mesures de similarité pour données binaires s'expriment en fonction de quatre quantités associées au couple (x, y) : les nombres d'attributs respectivement communs aux deux objets $a = |X \cap Y|$, présents dans x mais non dans y ou vice-versa, $b = |X - Y|$ et $c = |Y - X|$, et enfin présents ni dans x ni dans y, $d = |\bar{X} \cap \bar{Y}|$.

Les mesures principales sont rappelées dans le tableau 1, en distinguant deux catégories : les mesures dites *de type I*, qui correspondent aux quatre premières du tableau, sont indépendantes des attributs absents des deux objets, d, alors que les mesures dites *de type II* en dépendent. Aussi, pour ces dernières, la taille de l'univers dans lequel les données sont décrites influence la valeur de similarité.

On peut observer que les deux premières mesures, Jaccard et Dice, suivent le même schéma général, proposé par Tversky (1977), qui définit des mesures de la forme $\text{Tve}_{\alpha,\beta}(x, y) = a/(a + \alpha b + \beta c)$. Elles correspondent aux cas particuliers où $\alpha = \beta$ = 1 et 1/2 respectivement.

	Mesure de similarité	*Notation*	*Définition*
Type I	Jaccard	Jac	$\frac{a}{a+b+c}$
	Dice	Dic	$\frac{2a}{2a+b+c}$
	Kulczynski 2	Kul	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$
	Ochiai	Och	$\frac{a}{\sqrt{a+b}\sqrt{a+c}}$
Type II	Rogers & Tanimoto	RT	$\frac{a+d}{a+2(b+c)+d}$
	Russel & Rao	RR	$\frac{a}{a+b+c+d}$
	Simple Matching	SM	$\frac{a+d}{a+b+c+d}$
	Sokal & Sneath 1	$SS1$	$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$
	Yule Q	YuQ	$\frac{ad}{ad+bc}$
	Yule Y	YuY	$\frac{\sqrt{ad}}{\sqrt{ad}+\sqrt{bc}}$

TAB. 1 – *Mesures de similarité classiques pour données binaires, normalisées dans* $[0,1]$ *(les définitions peuvent donc légèrement différer des définitions usuelles).*

3.3 Cas des données numériques

3.3.1 Corpus de test

Comme pour les données binaires, les données numériques réelles proviennent du corpus du Challenge ImageCLEF (2008). Elles utilisent une représentation qui exploite uniquement les images elles-mêmes et non leurs annotations. Plus exactement, les images sont décrites par leurs histogrammes HSV, en utilisant $q = 6 \times 2 \times 2 = 24$ *bins*.

La génération des données artificielles est réalisée selon le même principe que dans le cas des données artificielles binaires. Elle est effectuée selon une distribution uniforme sur $[0, 100]^{24}$.

3.3.2 Mesures considérées

Les comparaisons de données numériques sont basées sur des distances ou des produits scalaires. Nous abordons successivement ces deux catégories, en notant les données, vecteurs de $\mathcal{X} = \mathbb{R}^q$, sous la forme $x = (x_i)_{i=1..q}$.

Distances Parmi les nombreuses mesures de distance existantes, les plus fréquentes sont la distance euclidienne et la distance de Manhattan, qui sont des cas particuliers de la distance de Minkowski définie de façon générale comme

$$d(x,y) = \left(\sum_{i=1}^{q} |x_i - y_i|^\gamma\right)^{1/\gamma}$$

	Mesure	*Notation*	*Définition*
Distances	euclidienne	L2	$d_e(x,y) = \sqrt{\sum_{i=1}^{q}(x_i - y_i)^2}$
	Manhattan	L1	$d_M(x,y) = \sum_{i=1}^{q} \lvert x_i - y_i \rvert$
Noyaux	produit scalaire euclidien	PSE	$ke(x,y) = {}^t xy$
	noyau gaussien	NGσ	$kg_\sigma = \exp\left(-\frac{\lVert x-y \rVert^2}{2\sigma^2}\right)$
	noyau polynomial	NPγ	$kp_{\gamma,l} = (\langle x,y \rangle + l)^\gamma$
Normalisation	Noyau normalisé	kN	$\tilde{k}(x,y) = \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$

TAB. 2 – *Mesures de comparaison classiques pour données numériques.*

en notant γ un réel positif. La distance euclidienne correspond à $\gamma = 2$, la distance de Manhattan à $\gamma = 1$. La valeur de γ détermine les propriétés de la distance. Ainsi, les distances de Manhattan ou de Tchebychev (cas limite où $\gamma \to \infty$) ne sont pas dérivables ; la distance de Manhattan est plus robuste que la distance euclidienne, car elle augmente moins rapidement et prend des valeurs plus faibles quand la base de données contient des données aberrantes. Ces propriétés peuvent servir de critères pour la sélection d'une distance plutôt qu'une autre.

Comme souligné précédemment, une mesure de distance a une sémantique opposée à celle d'une mesure de similarité, et un renversement d'échelle est nécessaire pour passer de l'une à l'autre et permettre une comparaison. Celle-ci se fait par l'application d'une fonction décroissante, comme par exemple une fonction linéaire, gaussienne ou de Cauchy, qui permettent de contrôler la vitesse de décroissance et le pouvoir de discrimination (Lesot et al., 2009). Le choix d'une telle fonction n'influence cependant pas l'ordre au-delà de l'inversion globale d'échelle, c'est-à-dire ne produit pas d'inversion locale entre paires d'objets. Aussi, il est hors du cadre de l'étude considérée ici, basée sur l'ordre induit ; les expériences sont réalisées en considérant l'ordre induit par $-d$. Il en va de même des opérations de normalisation parfois appliquées pour garantir des valeurs dans un intervalle fixé, le plus souvent $[0, 1]$.

Produits scalaires Depuis le développement des méthodes d'apprentissage à noyau, les données numériques sont très fréquemment comparées en utilisant un produit scalaire, que ce soit le produit scalaire euclidien ou un noyau, c'est-à-dire un produit scalaire appliqué à des données implicitement transformées par une fonction non linéaire. Les plus utilisés sont indiqués dans le tableau 2.

Comme souligné précédemment, à l'exception du noyau gaussien, un produit scalaire ne correspond pas à une mesure de similarité au sens classique, car il ne vérifie pas la propriété de maximalité : on a par exemple $k(x, 2x) > k(x, x)$. Pour obtenir cette propriété, il est

	Jac	Kul2	Och	RT	RR	SM	SS1	YuQ	YuY	Aléatoire
Dic	1	0.97	0.99	0.87	0.89	0.87	0.87	0.86	0.86	0.50
Jac		0.97	0.99	0.87	0.89	0.87	0.87	0.86	0.86	0.50
Kul2			0.98	0.88	0.88	0.88	0.88	0.88	0.88	0.50
Och				0.88	0.89	0.88	0.88	0.87	0.87	0.50
RT					0.76	1	1	0.90	0.90	0.50
RR						0.76	0.76	0.77	0.77	0.50
SM							1	0.90	0.90	0.50
SS1								0.90	0.90	0.50
YuQ									1	0.50
YuY										0.50

TAB. 3 – *Degrés d'équivalence calculés sur les listes entières pour les données binaires artificielles.*

nécessaire d'effectuer une normalisation, comme indiqué dans la dernière ligne du tableau 2. Les variantes normalisées du produit scalaire euclidien et du noyau polynomial sont notées PSEN et NPN respectivement. Un tel produit scalaire normalisé mesure alors la similarité entre vecteurs en fonction de l'angle qu'ils forment : il est maximal pour des vecteurs de même direction, et minimal pour des vecteurs de direction opposée. La comparaison qu'il effectue a alors une sémantique essentiellement différente de celle d'une mesure de distance, y compris de la mesure de distance qu'il induit, définie par $d(x, y) = \sqrt{k(x - y, x - y)}$.

4 Résultats de quasi-équivalence pour les mesures de comparaison de données binaires

Nous dressons dans cette section une synthèse des résultats obtenus sur l'équivalence et la quasi-équivalence des mesures de comparaison de données binaires en utilisant le protocole expérimental décrit dans la section 3.1. Nous examinons tout d'abord la comparaison des listes induites dans leur totalité, puis la comparaison des listes restreintes aux éléments dont le rang est inférieur ou égal à un paramètre k, avec $k = 100$ et $k = 10$.

4.1 Comparaison des listes entières

Le tableau 3 contient les degrés d'équivalence calculés sur les listes entières pour les données binaires artificielles.

Mesures équivalentes On peut observer que trois groupes de mesures ont des degrés d'équivalence valant 1, indiquant des mesures équivalentes : Jaccard et Dice ; Rogers & Tanimoto, Simple Matching et Sokal & Sneath 1 ; Yule Q et Yule Y. Ces résultats sont en accord avec les résultats analytiques connus : Batagelj et Bren (1995); Omhover et al. (2006), en utilisant la définition fonctionnelle de l'équivalence et en identifiant les fonctions impliquées, ont établi les classes de mesures de similarité équivalentes suivantes

Degrés d'équivalence de mesures de comparaison

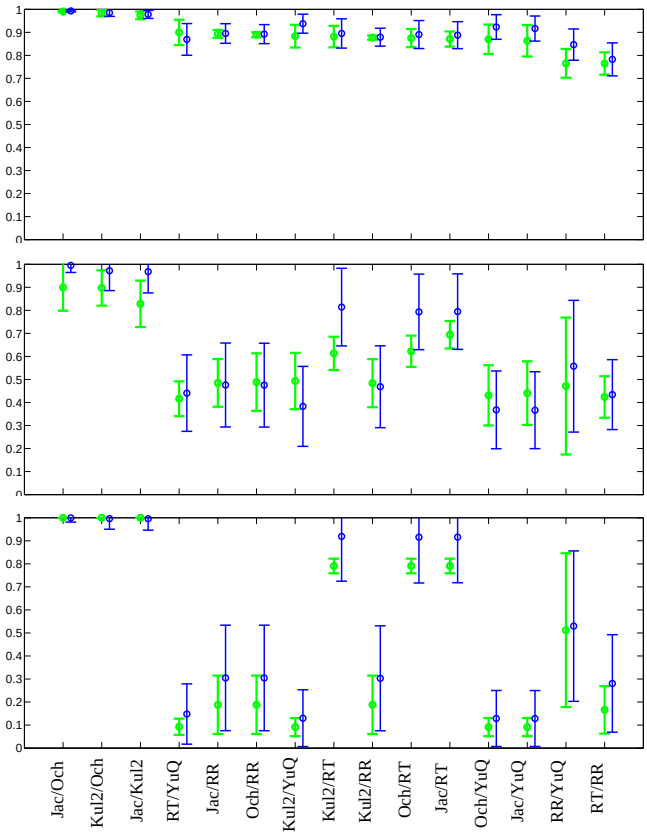

FIG. 3 – *Degrés d'équivalence et leurs écarts-types pour les données binaires : listes entières (en haut), listes tronquées pour k = 100 (au milieu) et k = 10 (en bas). Pour chaque paire de mesures, la barre verte à gauche (resp. bleue, à droite) correspond aux données artificielles (resp. réelles).*

- {Jaccard, Dice, les mesures de Tversky symétriques $\text{Tve}_{\alpha,\alpha}$},
- {Rogers & Tanimoto, Simple Matching, Sokal & Sneath 1},
- {Yule Q, Yule Y}
- chaque mesure restante constitue une classe réduite à elle-même.

Pour les mesures de Tversky, il a été démontré plus généralement (Omhover et al., 2006) que deux mesures de paramètres (α, β) et (α', β') sont équivalentes si et seulement si $\alpha/\beta = \alpha'/\beta'$.

Mesures non équivalentes Les degrés différents de 1 quant à eux renseignent sur les mesures non équivalentes. On peut tout d'abord noter que la mesure aléatoire a un degré d'équivalence

de 0.5 avec toutes les autres mesures, ce qui constitue également un résultats attendu : pour toute paire de données, elle a une chance sur deux d'avoir le même classement que la mesure à laquelle elle est comparée ; aussi, en moyenne, elle classe différemment la moitié des paires d'objets.

On peut également observer que tous les degrés sont élevés : hormis la mesure aléatoire, le degré minimal vaut 0.76, c'est-à-dire la proportion d'inversions ne dépasse pas 24%. Ceci signifie que les ordres induits ne diffèrent que peu, et que les mesures sont redondantes.

De plus, on observe que certaines mesures, bien que non équivalentes au sens strict, sont équivalentes à un très haut degré, supérieur à 0.97 : Jaccard, Ochiai et Kulczynski 2, qui correspondent aux mesures de type I, ne diffèrent donc que très peu. Elles peuvent donc être considérées comme quasi-équivalentes.

Le graphique supérieur de figure 3 illustre ces degrés avec leurs écarts-types, calculés lorsque la donnée qui sert de requête varie. Elle indique également les valeurs obtenues lors du calcul du degré d'équivalence sur les données réelles, par la barre en bleu à droite. Elle représente les paires de mesures dans l'ordre décroissant de leurs degrés d'équivalence ; pour améliorer la lisibilité, un seul représentant de chaque classe d'équivalence est mentionné, et la mesure aléatoire est omise.

Les barres d'erreur représentant les écarts-types montrent qu'il n'y a pas de différence significative entre les résultats obtenus sur les données artificielles et les données réelles. Tous les commentaires sur les mesures sont donc également valables pour ces dernières.

Le graphique souligne également les différences entre les deux types de mesures, comme mentionné précédemment : alors que les mesures de type I sont largement équivalentes entre elles, la disparité entre mesures de type II est nettement plus importante. Ces dernières ne se ressemblent pas plus entre elles qu'elles ne ressemblent aux mesures de type I, ce qui souligne l'hétérogénéité de cette catégorie.

4.2 Comparaison des listes tronquées

Les deux graphiques inférieurs de la figure 3 montrent les degrés d'équivalence obtenus en considérant les classements tronqués, c'est-à-dire restreints aux données de rang inférieur ou égal à k, pour $k = 100$ et $k = 10$ respectivement. Ils utilisent le même axe des abscisses que le graphique obtenu sur les listes entières afin de souligner les différences qui surviennent quand on considère les listes tronquées.

Faibles degrés d'équivalence On peut tout d'abord observer que les degrés sont globalement plus faibles que pour la comparaison des listes entières : le minimum est 0.42 pour $k = 100$, 0.09 pour $k = 10$, ce qui indique des différences majeures dans les classements fournis par les mesures. Le degré d'équivalence de la mesure aléatoire avec toute autre mesure (non montré sur les graphes) est inférieur à 0.1 : l'ordre qu'elle induit n'a presque rien en commun avec chacune des autres mesures.

Cette diminution indique que l'accord global observé lors de la comparaison des listes entières est principalement dû aux dernières données classées, et que de fortes différences existent en tête de liste. Elle souligne donc l'importance de l'étude des positions des inversions, au-delà de celle de leur nombre, en particulier quand l'objectif est de choisir des mesures non équivalentes dans un cadre de recherche d'information.

Néanmoins, cette diminution ne se produit pas pour toutes les mesures : les paires comparant des mesures de type I entre elles ou avec la mesure Rogers & Tanimoto (RT) sont stables lorsqu'on passe de l'étude des listes entières aux listes tronquées. En raison de ce comportement, RT apparaît, malgré son appartenance à la catégorie des mesures de type II, comme plus proche des mesures de type I que des mesures de sa propre catégorie. Ces mesures peuvent être considérées comme équivalentes même pour des ordres restreints, et donc redondantes dans le cadre d'application de recherche d'information.

Evolution des écarts-types Une autre différence observée lors de la restriction aux listes tronquées provient des écart-types : on peut observer que leurs valeurs augmentent considérablement. De plus, ils prennent en général des valeurs plus élevées sur les données réelles que sur les données artificielles. Ce comportement peut être dû à la distribution régulière des données artificielles, qui garantit une certaine indépendance par rapport à la donnée requête. Au contraire, les données réelles suivent une distribution à densité variable, et la donnée requête peut avoir des effets différents, selon qu'elle appartient à une région dense ou non. Néanmoins, comme pour la comparaison de listes entières, et à l'exception de la mesure Rogers & Tanimoto, on n'observe pas de différence significative entre les données artificielles et réelles.

Mesures isolées : Yule Q et Russel & Rao Il apparaît que les mesures de Yule Q et Russel & Rao deviennent les plus isolées, très éloignées des autres mesures. Pour Yule Q, ceci peut être expliqué par le fait qu'elle prend très souvent la valeur 1. En effet, celle-ci est obtenue pour toutes les paires (x, y) telles que $b = 0$ ou $c = 0$. Aussi, l'ensemble de données de rang inférieur ou égal à k est bien plus grand que pour les autres mesures, conduisant à de très nombreuses paires manquantes lors de la comparaison des ordres induits.

Le comportement de Russel & Rao (RR) peut être expliqué de façon similaire : cette mesure prend seulement $q + 1 = 12$ valeurs distinctes dans un univers de dimension q. Aussi, les listes tronquées contiennent la totalité des données, même pour de faibles valeurs de k, conduisant également à de nombreuses paires manquantes lors de la comparaison avec d'autres ordres.

Liste de mesures candidates Cette étude permet d'identifier un ensemble représentatif et complet de mesures candidates, permettant à la fois de couvrir les différentes comportements possibles sans introduire de redondance : il doit contenir une seule mesure de type I, qui suffit à représenter toutes cette catégorie. En ce qui concerne les mesures de type II, il doit comporter une mesure de chacune des classes d'équivalence. Toutefois, si une variabilité de moins de 0.8 n'est pas considérée comme significative, la mesure Rogers & Tanimoto peut être éliminée, du fait de sa ressemblance élevée aux mesures de type I.

5 Résultats de quasi-équivalence pour les mesures de comparaison de données numériques

Nous examinons dans cette section les équivalences et quasi-équivalences qui peuvent être établies pour les mesures de comparaison pour données numériques indiquées dans le tableau 2. Nous les étudions d'abord d'un point de vue analytique, puis expérimentalement, sur des données artificielles et réelles.

5.1 Résultats analytiques

L'étude analytique de l'équivalence entre mesures est basée sur sa définition fonctionnelle et l'identification des fonctions de passage entre les mesures. Elle permet d'identifier deux classes d'équivalence.

La première regroupe, de façon évidente, les noyaux gaussiens avec la distance euclidienne. En effet, on a $kg_\sigma = f \circ d_e$ avec $f : x \mapsto \exp(-x^2/(2\sigma^2))$ qui est décroissante. Tous les noyaux gaussiens sont donc équivalents, à la distance euclidienne et entre eux. Ceci implique en particulier que toutes les valeurs de σ conduisent toujours au même ordre des données quand elles sont comparées à une requête.

La seconde classe, qui regroupe le produit scalaire euclidien et les noyaux polynomiaux, est définie à une translation des données près : pour les valeurs paires de γ, la fonction $g(x) = (x + l)^\gamma$, telle que $kp_{\gamma,l} = g \circ k_e$, est croissante seulement à la condition que $x \geq -l$. Si on note α la valeur telle que $\forall x \forall i,\ x_i + \alpha \geq 0$ et e le vecteur tel que $\forall i,\ e_i = \alpha$, la translation par e garantit que $\forall x \forall i,\ x_i \geq 0$ et donc $\langle x, y\rangle = \sum_i x_i y_i \geq 0 > -l$. On peut souligner que dans un cadre de classification, la valeur de l est sans importance, car elle pondère les attributs dans l'espace des description, et est contrebalancée par les poids appris par le classifieur.

Ces deux classes sont illustrées sur la figure 1 : on observe que le noyau euclidien et la distance euclidienne d'une part, le produit scalaire euclidien et le noyau polynomial d'autre part, conduisent aux mêmes listes d'images.

Il faut noter que dans le cas où les données sont normalisées, de telle sorte que pour tout x, $\|x\| = 1$, les deux classes d'équivalence précédentes n'en forment qu'une : on a $d_e = h \circ ke$ avec $h(x) = \sqrt{2(1 - x)}$ qui est strictement décroissante.

5.2 Comparaison des listes entières

Dans le cadre de l'étude expérimentale, nous considérons les mesures rappelées dans le tableau 2 avec les paramètres suivants ; nous considérons la distance euclidienne, notée L2, et la distance de Manhattan, notée L1, le produit scalaire euclidien (PSE) et sa forme normalisée (PSEN), le noyau gaussien pour $\sigma = 50$ (NG50) et $\sigma = 100$ (NG100), le noyau polynomial de degré 3 pour $l = 2000$ (NP3) et sa normalisation (NP3N). Les valeurs des paramètres σ et l pour les noyaux gaussiens et polynomiaux ont été choisis selon les propriétés des données. Enfin nous incluons la mesure aléatoire, à titre de comparaison.

Le tableau 4 contient les degrés d'équivalence entre ces 9 mesures, calculés sur les listes entières pour les données numériques artificielles.

Comme dans le cas binaire, et pour la même raison, on observe que la mesure aléatoire a un degré de 0.5 avec toutes les autres mesures. Les degrés valant 1 sont en accord avec les résultats théoriques et indiquent les deux classes attendues. A nouveau, toutes les mesures ont un assez haut niveau d'accord, puisque la valeur minimale des degrés d'équivalence est de 0.63, c'est-à-dire la proportion maximale d'inversion est de 37% dans le pire cas, obtenu pour les noyaux polynomiaux et gaussiens.

Le degré d'équivalence élevé entre L2 et NP3N, de 0.97, ne correspond pas à un résultat connu théoriquement. Il peut être expliqué en considérant les lignes de niveaux de ces mesures, représentées pour des données 2D sur la figure 4 : même si elles diffèrent localement, leur forme générale est similaire et les ordres induits concordent globalement.

Degrés d'équivalence de mesures de comparaison

	L2	PSE	PSEN	NG50	NG100	NP3	NP3N	Aléatoire
L1	0.90	0.63	0.84	0.90	0.90	0.63	0.89	0.50
L2		0.63	0.87	1	1	0.63	0.97	0.50
PSE			0.76	0.63	0.63	1	0.66	0.50
PSEN				0.87	0.87	0.76	0.90	0.50
NG50					1	0.63	0.97	0.50
NG100						0.63	0.97	0.50
NP3							0.66	0.50
NP3N								0.50

TAB. 4 – *Degrés d'équivalence calculés sur les listes entières pour les données numériques artificielles.*

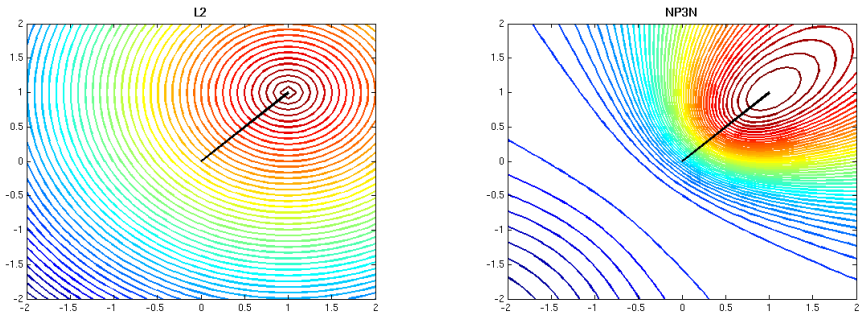

FIG. 4 – *Lignes de niveaux pour L2 et NP3N, avec pour référence le point (1,1).*

Le graphique supérieur de la figure 5 offre une représentation graphique des degrés d'équivalence présentés dans le tableau 4 et présente les résultats obtenus pour les données réelles, selon les mêmes principes que la figure 3 (résultats moyens avec écart-type quand la donnée requête varie, à gauche en vert les données artificielles, à droite en bleu les données réelles, un seul représentant par classe d'équivalence de mesures). Il montre qu'il existe une différence importante, qui conduit à un ordre différent des paires de mesures d'après leurs degrés d'équivalence. On peut l'expliquer par la particularité des données réelles : comme elles correspondent à des histogrammes de répartition, leur norme L1 est constante. Cette structure spécifique des données a des conséquences sur leurs degrés d'équivalence.

5.3 Comparaison des listes tronquées

Lorsqu'on considère les classements restreints à leurs premiers éléments (graphes inférieurs de la figure 5), on observe que la différence entre les deux types de données, réelles et artificielles, devient moins marquée quand k diminue. Les écart-types augmentent, soulignant l'influence de la donnée requête, en particulier sur les têtes de liste.

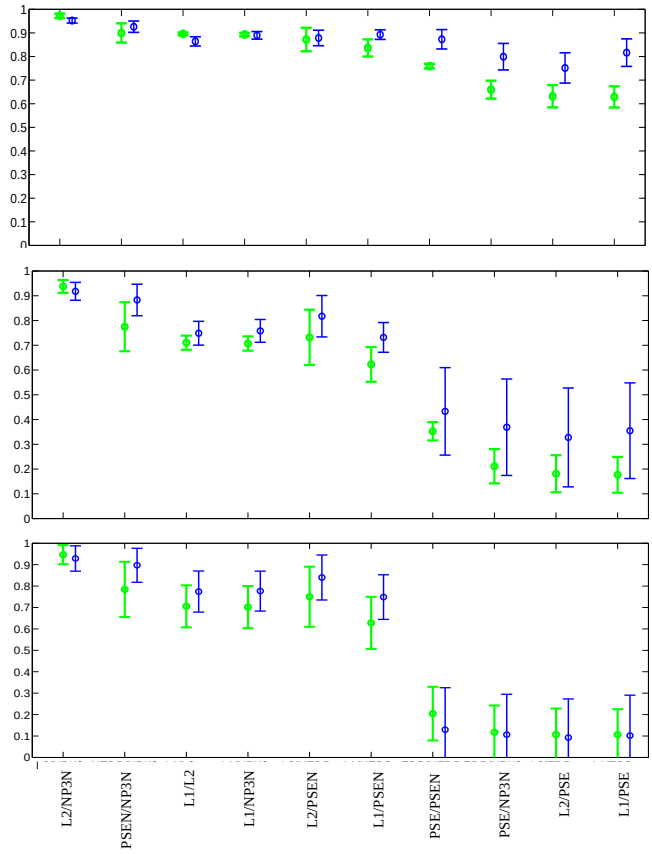

FIG. 5 – *Degrés d'équivalence et leurs écarts-types pour les données numériques : (en haut) listes entières, (au milieu) listes tronquées, k = 100, (en bas) listes tronquées, k = 10. Pour chaque paire de mesures, la barre verte à gauche (resp. bleue, à droite) correspond aux données artificielles (resp. réelles).*

Niveaux d'équivalence Bien que les degrés d'équivalence diminuent fortement par rapport aux cas des listes entières, l'ordre des paires de mesures n'est pas modifié significativement. Trois niveaux d'équivalence peuvent être distingués, en particulier pour $k = 10$. Le niveau le plus élevé est atteint par la paire L2/NP3N, ce qui signifie que leur accord élevé est valable pour les valeurs de similarité les plus élevées. Elles peuvent donc être considérées comme quasi-équivalentes et redondantes.

Une mesure isolée : le produit scalaire euclidien Au contraire, PSE conduit à des valeurs très faibles, quelle que soit la mesure à laquelle il est comparé, qui soulignent sa non redondance : il fournit des listes ordonnées significativement différentes de toutes les autres mesures, et reste candidat dans la liste des mesures à considérer lors du choix d'une mesure dans la conception d'un système.

Liste de mesures candidates Comme dans le cas des mesures pour données binaires, cette étude permet d'identifier un ensemble de mesures à la fois complémentaires et non redondantes qui constituent des mesures candidates pour la mise au point d'une système de recherche d'information ou d'apprentissage où seul l'ordre induit par les mesures importe. Ces mesures sont le produit scalaire euclidien, qui est nettement différent de toutes les autres mesures, puis la distance euclidienne (ou le noyau polynomial normalisé), la distance de Manhattan et le produit scalaire euclidien normalisé.

6 Conclusion

Nous avons comparé les mesures de similarité pour deux types de données, quantifiant leur proximité et redondance possible lorsqu'on étudie les classements qu'elles induisent, et en particulier la restriction de ces derniers aux têtes de listes. Cette analyse, qui repose sur la définition de degrés d'équivalence basés sur le coefficient de Kendall généralisé, se place dans le cadre des systèmes de recherche d'information. Les expériences réalisées sur des données artificielles et réelles montrent des propriétés de stabilité des comportements des mesures de comparaison, mais également des différences qui soulignent que les degrés d'équivalence dépendent, dans une certaine mesure, des données considérées.

Les perspectives de ce travail visent à examiner les diverses configurations et propriétés de données qui peuvent conduire à des équivalences en pratique, permettant d'approfondir l'aide au choix des mesures de comparaison. Lerman (1967) a réalisé une telle étude dans le cas des données binaires, montrant que si toutes les données ont le même nombre d'attributs présents, c'est-à-dire si $\exists \eta / \forall x \in \mathcal{D}\ |X| = \eta$ alors toutes les mesures de similarité sont équivalentes sur $\mathcal{D}$. L'objectif est d'étendre cette étude aux degrés d'équivalence ainsi qu'aux données numériques.

Références

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

Baulieu, F. B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification 6*, 233–246.

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

Challenge ImageCLEF (2008). http ://www.imageclef.org.

Fagin, R., R. Kumar, M. Mahdian, D. Sivakumar, et E. Vee (2004). Comparing and aggregating rankings with ties. In *Symp. on Princ. of Database Sys.*, pp. 47–58.

Fagin, R., R. Kumar, et D. Sivakumar (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics 17*(1), 134–160.

Janowitz, M. F. (1979). Monotone equivariant cluster analysis. *SIAM J. Appl. Math. 37*, 148–165.

Lerman, I. C. (1967). Indice de similarité et préordonnance associée. In *Séminaire sur les ordres totaux finis*, Aix-en-Provence, pp. 233–243.

Lesot, M.-J., M. Rifqi, et H. Benhadda (2009). Similarity measures for binary and numerical data : a survey. *Intern. J. of Knowledge Engineering and Soft Data Paradigms (KESDP) 1*(1), 63–84.

Omhover, J.-F., M. Rifqi, et M. Detyniecki (2006). Ranking invariance based on similarity measures in document retrieval. In *Adaptive Multimedia Retrieval AMR'05*, pp. 55–64. Springer LNCS.

Rifqi, M., V. Berger, et B. Bouchon-Meunier (2000). Discrimination power of measures of comparison. *Fuzzy sets and systems 110*, 189–196.

Rifqi, M., M. Detyniecki, et B. Bouchon-Meunier (2003). Discrimination power of measures of resemblance. In *Proceedings of the International Fuzzy Systems Association World Congress - IFSA'03*, pp. 503–510.

Rifqi, M., M.-J. Lesot, et M. Detyniecki (2008). Fuzzy order-equivalence for similarity measures. In *Proc. of NAFIPS 2008*.

Tversky, A. (1977). Features of similarity. *Psychological Review 84*, 327–352.

Summary

In order to help choosing a measure to compare data, which constitutes an essential step in the conception of systems for data mining, machine learning or information retrieval, we compare the most common measures according to the order they induce and quantify their agreement by a degree of equivalence. We propose a systematic study of measures applied to two data types, namely binary and numerical data, considering the major similarity measures, distances and dot products. We establish their equivalence degrees both on artificial and real data sets, et identify equivalent and quasi-equivalent measures that can be considered as redundant in the framework of information retrieval.