Comparaison de distances et noyaux classiques par degré d'équivalence des ordres induits

Marie-Jeanne Lesot, Maria Rifqi, Marcin Detyniecki

Université Pierre et Marie Curie - Paris 6, CNRS UMR 7606, LIP6,
104 avenue du Président Kennedy, F-75016 Paris, France
{marie-jeanne.lesot,maria.rifqi,marcin.detyniecki}@lip6.fr

Résumé. Le choix d'une mesure pour comparer les données est au cœur des tâches de recherche d'information et d'apprentissage automatique. Nous considérons ici ce problème dans le cas où seul l'ordre induit par la mesure importe, et non les valeurs numériques qu'elle fournit : cette situation est caractéristique des moteurs de recherche de documents par exemple. Nous étudions dans ce cadre les mesures de comparaison classiques pour données numériques, telles que les distances et les noyaux les plus courants. Nous identifions les mesures équivalentes, qui induisent toujours le même ordre ; pour les mesures non équivalentes, nous quantifions leur désaccord par des degrés d'équivalence basés sur le coefficient de Kendall généralisé. Nous étudions les équivalences et quasi-équivalences à la fois sur les plans théorique et expérimental.

1 Introduction

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 un document candidat et la requête de l'utilisateur. Le choix de la mesure de similarité, ou plus généralement de la mesure de comparaison, est alors au cœur de la conception du système. Pour de telles applications, ce sont les ordres induits par les mesures de comparaison qui importent et non les valeurs numériques qu'elles prennent : 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 de conserver des mesures qui donnent le même classement des données.

La notion d'équivalence entre mesures de comparaison en terme d'ordre a été introduite initialement pour les mesures de similarité pour données ensemblistes (Lerman, 1967; Baulieu, 1989; Batagelj et Bren, 1995; Omhover et al., 2006). Elle a été raffinée par la définition de degrés d'équivalence permettant d'examiner plus finement les écarts entre mesures non équivalentes, en quantifiant le désaccord entre les ordres qu'elles induisent (Rifqi et al., 2008).

Dans cet article, nous considérons la problématique de l'équivalence de mesures de comparaison dans le cas des données numériques, en examinant les mesures classiques, incluant les distances et les noyaux les plus courants. Dans la section 2 nous rappelons les définitions de l'équivalence et des degrés d'équivalence. La section 3 expose les résultats obtenus pour les mesures classiques pour données numériques, à la fois sur les plans théorique et expérimental, après application des mesures de comparaison à une base d'images.

2 Degrés d'équivalence entre mesures de comparaison

Equivalence de mesures de comparaison Une mesure de comparaison est une fonction qui associe à toute paire d'objets une valeur quantifiant leur ressemblance ou, de façon duale, leur dissimilarité. La comparaison théorique de ces mesures (Lerman, 1967; Baulieu, 1989; Batagelj et Bren, 1995; Omhover et al., 2006), pour des données ensemblistes, a conduit à définir la notion d'équivalence : deux mesures m_1 et m_2 sont dites *équivalentes* si et seulement si elles induisent le même ordre, c'est-à-dire si et seulement si $\forall x, y, z, t$, on a $m_1(x, y) < m_1(z, t) \Leftrightarrow m_2(x, y) < m_2(z, t)$ et $m_1(x, y) = m_1(z, t) \Leftrightarrow m_2(x, y) = m_2(z, t)$.

De façon équivalente (Batagelj et Bren, 1995; Omhover et al., 2006), deux mesures m_1 et m_2 sont équivalentes si et seulement si $\exists f : Im(m_1) \rightarrow Im(m_2)$ 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.

Degrés d'équivalence Rifqi et al. (2008) ont proposé de quantifier le désaccord entre mesures non équivalentes, par des *degrés d'équivalence*. En effet, deux mesures qui ne conduisent qu'à quelques inversions sont "plus fortement" équivalentes que si elles induisent des ordres opposés ; elles le sont d'autant plus que les inversions se produisent pour des valeurs de similarité faibles : pour les moteurs de recherche par exemple, le plus souvent, seuls les premiers résultats sont pris en compte, des inversions en fin de classement ne sont pas même remarquées.

Aussi, Rifqi et al. (2008) ont proposé de tenir compte du nombre d'inversions et de leurs positions, en utilisant le coefficient de Kendall généralisé $K_{p,p'}$ (Fagin et al., 2003, 2004) : celui-ci associe à chaque paire d'objets (i, j) une pénalité $P(i, j)$ puis calcule la somme des pénalités rapportée au nombre de paires considérées. Quatre valeurs de pénalité sont distinguées, suivant que la paire est concordante ($P = 0$), discordante ($P = 1$), ex-aequo dans l'un des classements mais non dans l'autre ($P = p \in [0, 1]$), ou enfin présente dans l'une des listes mais non dans l'autre. Dans ce dernier cas, on distingue suivant que i et j sont tous deux absents ($P = p' \in [0, 1]$), ou qu'un seul l'est (la paire est traitée comme une paire normale).

Pour comparer deux mesures m_1 et m_2, on classe un ensemble de données $\mathcal{D}$ suivant leur similarité à une donnée de référence, puis on tronque le classement, pour ne considérer que les objets de rang inférieur à un paramètre k. En notant r^k cet ordre restreint, le degré d'équivalence est défini comme (Rifqi et al., 2008)

$$d^k_{\mathcal{D}}(m_1, m_2) = 1 - K_{0.5,1}(r^k_1, r^k_2)$$

Il vaut 1 pour des mesures équivalentes, et 0 des mesures induisant des classements opposés. On fixe $p = 0,5$ et $p' = 1$: en choisissant un ordre strict pour départager des ex-aequo, on a une chance sur deux d'obtenir l'ordre de l'autre liste ; de plus, une paire manquante indique une différence majeure entre deux listes, et peut être pénalisée autant qu'une paire discordante.

3 Equivalence des mesures pour données numériques

Nous rappelons ici les mesures de comparaison classiques pour données numériques, puis présentons les résultats d'équivalence et de quasi-équivalence obtenus théoriquement et expérimentalement : ils montrent qu'une équivalence théorique peut conduire à des différences en pratique, bien que de faible importance, et que, réciproquement, certaines propriétés des données peuvent conduire à des proximités qui ne sont pas vraies dans le cas général.

3.1 Mesures de comparaison pour données numériques

Les comparaisons de données numériques sont basées sur des distances ou des produits scalaires (Lesot et al., 2008). Les distances possèdent des propriétés de positivité, symétrie, et minimalité ($d(x,y) = 0 \Leftrightarrow x = y$), équivalentes aux propriétés des mesures pour données ensemblistes. De plus, elles vérifient 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 les plus courants sont le produit scalaire euclidien $ke = \langle x,y \rangle$, les noyaux gaussien $kg_\sigma = \exp(-\|x-y\|^2/(2\sigma^2))$ et polynomial $kp_{\gamma,l} = (\langle x,y \rangle + l)^\gamma$. A l'exception du noyau gaussien, ils ne correspondent pas à une similarité classique, car ils ne vérifient pas la propriété de maximalité : on a par exemple $k(x,2x) > k(x,x)$. Pour obtenir cette propriété, il est nécessaire de les normaliser, en définissant $\bar{k}(x,y) = k(x,y)/\sqrt{k(x,x)k(y,y)}$. La similarité dépend alors uniquement de l'angle formé par les 2 vecteurs.

3.2 Etude analytique

En utilisant la définition fonctionnelle de l'équivalence, deux classes d'équivalence peuvent être distinguées immédiatement : la première regroupe les noyaux polynomiaux et le produit scalaire euclidien. En effet, pour tous l et γ positifs, on a $kp_{\gamma,l} = f \circ ke$ avec $f(x) = (x+l)^\gamma$ qui est strictement croissante. On a donc aussi équivalence des noyaux polynomiaux entre eux. La seconde classe regroupe les noyaux gaussiens et la distance euclidienne[1] d, car $kg_\sigma = g \circ (-d)$ avec $g(x) = \exp(-x^2/(2\sigma^2))$ qui est strictement décroissante.

Dans le cas où les données sont normalisées, ces deux classes n'en forment qu'une car on a $-d(x,y) = f(\langle x,y \rangle)$ avec $f(x) = -\sqrt{2(1-x)}$ qui est strictement croissante.

On ne peut établir d'autres équivalences, notamment pour les produits scalaires avant et après normalisation, ni entre les différentes distances de Minkowski. Les sémantiques de ces mesures diffèrent considérablement, et si une certaine concordance entre les ordres qu'elles induisent est attendue, un nombre non négligeable d'inversions est également à prévoir.

3.3 Etude expérimentale

Protocole On considère les 9 mesures indiquées dans le tableau 1 qui donne les degrés d'équivalence calculés sur une base de données contenant 18724 images issues des pages tourisme du site web Wikipedia français (Ah-Pine et Renders, 2007) : les classements des images selon leur similarité à une image requête choisie aléatoirement sont comparés, dans leur totalité ou après restriction aux images de rang inférieur à $k = 10$. Les résultats ont été moyennés sur 10 tests ; l'écart-type est très faible dans tous les cas, et reste toujours inférieur à 0,06.

Afin de mieux exploiter ces degrés, un clustering hiérarchique avec chaînage complet est appliqué, avec pour distance entre mesures $1 - d_k(m_1, m_2)$ (cf fig. 1) : ce chaînage fournit une borne supérieure du degré d'équivalence au sein de chaque cluster, qui renseigne sur les degrés avec lesquels les mesures du cluster peuvent être considérées comme équivalentes. On peut alors identifier des classes de quasi-équivalence en définissant un seuil de coupure du dendrogramme en fonction de la proportion minimale de concordances que l'on souhaite observer.

[1] Dans le cas d'une distance, il faut considérer $-d$ et non d pour inverser l'échelle des valeurs et pouvoir comparer aux mesures de similarité.

	L2	PSE	PSEN	NG1	NG10	NP3	NP3N	Aléatoire
L1	0,81	0,6	0,58	0,8	0,81	0,6	0,82	0,5
L2		0,57	0,56	0,98	1	0,57	0,93	0,5
PSE			0,96	0,58	0,57	1	0,65	0,5
PSEN				0,56	0,56	0,96	0,63	0,5
NG1					0,98	0,58	0,91	0,5
NG10						0,57	0,93	0,5
NP3							0,65	0,5
NP3N								0,5
L1	0,53	0,2	0,52	0,53	0,53	0,21	0,57	0
L2		0,19	0,57	1	0,99	0,19	0,69	0
PSE			0,33	0,19	0,18	1	0,24	0
PSEN				0,57	0,58	0,33	0,78	0
NG1					0,99	0,19	0,69	0
NG10						0,18	0,69	0
NP3							0,24	0
NP3N								0

TAB. 1 – *Degrés d'équivalence, en haut, pour les listes ordonnées entières, en bas, pour les listes tronquées ($k = 10$). L1 et L2 désignent les distances de Manhattan et euclidienne, PSE et PSEN les produits scalaires ke et $\widetilde{ke}$, NG1, NG10, NP3 et NP3N les noyaux kg_1 kg_{10}, $kp_{3,1}$ et $\widetilde{kp}_{3,1}$ respectivement, et Aleatoire une mesure de référence produisant un ordre aléatoire.*

Comparaison des listes entières On observe dans la partie supérieure du tableau 1 deux paires de mesures équivalentes. La première, conformément au résultat attendu, groupe le produit scalaire euclidien et le noyau polynomial ; la seconde associe la distance euclidienne et le noyau gaussien pour $\sigma = 10$, mais exclut $\sigma = 1$ dont le degré est 0,98 et non 1. En effet, NG1 conduit à de nombreuses valeurs très proches de 0, qui sont traitées comme des ex-aequo.

La mesure aléatoire a un degré d'équivalence de 0,5 avec toutes les mesures : en moyenne, elle classe différemment la moitié des paires de données. Le fort degré d'équivalence de NP3N avec L2 ne correspond pas à un résultat analytiquement connu. Il peut être expliqué par les lignes de niveaux de ces mesures (figure omise pour des contraintes de place) : même si elles diffèrent localement, leur forme générale est similaire et les ordres induits concordent globalement. Le degré élevé de PSEN et PSE, est surprenant car L2 et PSE ont un degré plutôt faible, alors qu'une équivalence des trois mesures est attendue si les données sont normalisées (cf section 3.2). La proximité observée est probablement due à une configuration particulière des données étudiées ici, telle que celle illustrée et commentée pour des données 2D sur la figure 2.

Le dendrogramme (figure 1 gauche) conduit à distinguer 3 groupes : (i) la mesure aléatoire, (ii) le produit scalaire euclidien, normalisé ou non, et le noyau polynomial non normalisé, (iii) les distances L1 et L2, les noyaux polynomial normalisé et gaussiens. Aussi, pour ces données, dans le cas d'une recherche par similarité, il est inutile de proposer toutes les mesures à l'utilisateur, ou 5 mesures des 5 classes d'équivalence établies analytiquement : 2 suffisent (la mesure aléatoire étant évidemment exclue). Avec un seuil de concordance minimale de 90%, la coupure du dendrogramme à 0,1 indique que L1 devient une mesure à proposer.

Comparaison des listes tronquées Dans la partie inférieure du tableau 1, les degrés d'équivalence atteignent des valeurs très basses, qui indiquent que les mesures diffèrent pour les

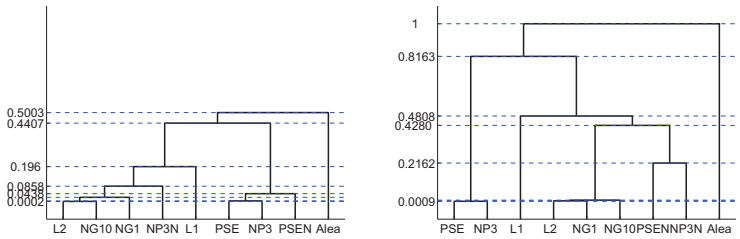


FIG. 1 – *Dendrogrammes basés sur les degrés d'équivalence : à gauche, pour les listes ordonnées entières, à droite pour les listes tronquées ($k = 10$).*

valeurs de similarité élevées : leur accord lors des comparaisons globales concerne principalement les données de classement inférieur. Aléatoire a toujours un degré 0 : la liste qu'elle induit n'a probablement aucune image commune avec les autres, et la totalité des paires comparées correspond à des paires manquantes. Le dendrogramme obtenu alors (figure 1 droite) diffère du précédent, principalement du fait de PSEN qui s'éloigne de PSE et se rapproche des distances. Les données de la figure 2 peuvent aussi expliquer cet éloignement : les données les plus similaires à la requête sont des éléments de C1, pour PSE comme pour PSEN. Les différences de normes dans C1 conduisent à des inversions, significatives proportionnellement au nombre de paires considérées, alors qu'elles ne l'étaient pas pour les listes entières.

Le dendrogramme conduit à distinguer 3 groupes : (i) la mesure aléatoire, (ii) le produit scalaire euclidien et le noyau polynomial de degré 3 non normalisés, (iii) les autres mesures, qui peuvent être décomposées en 3 : L1, le sous-groupe L2/NG1/NG10 et le sous-groupe PSEN/NP3N. Avec un seuil de concordance minimale de 90%, il faut distinguer 6 groupes, qui correspondent aux classes établies théoriquement.

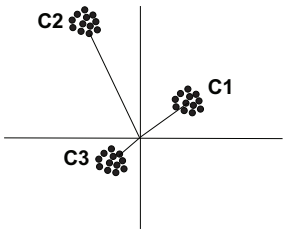
4 Conclusion

En exploitant les degrés d'équivalence, nous avons quantifié les variations observées entre les ordres induits par les mesures de comparaison classiques pour données numériques. Nous avons identifié analytiquement des classes d'équivalence, indiquant que l'on peut limiter le nombre de choix offerts à l'utilisateur d'un système de recherche par similarité. Des tests expérimentaux réalisés sur une base d'images ont de plus montré que ces résultats théoriques peuvent être complétés selon les propriétés des bases de données utilisées. Certaines remarques sont générales, portant par exemple sur la normalisation des données, d'autres sont plus spécifiques et dépendent de la structure des données en clusters et de leurs positions relatives.

Les perspectives de ce travail visent à examiner d'un point de vue théorique ces 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.

Remerciements

Nous remercions Xerox d'avoir mis à notre disposition les images Wikipedia indexées.



Pour une donnée requête appartenant à C1
– PSEN classe comme plus similaires d'abord les éléments de C1, puis ceux de C2 et enfin ceux de C3
– PSE produit le même classement : les éléments de C3 étant à l'opposé de la requête, ils conduisent à des similarité inférieures à celles des éléments de C2 ; les variations de normes à l'intérieur de chaque cluster étant petites, l'influence de la norme reste négligeable.
– pour L2, les données du cluster C3 sont plus proches de la requête que les données du cluster C2.

FIG. 2 – *Configuration telle que PSE soit fortement équivalent à PSEN et faiblement à L2.*

Références

Ah-Pine, J. et J.-M. Renders (2007). Mise en place du corpus pour l'évaluation de la recherche multi-média texte/image. Technical report, Rapport Infomagic D1.4-21.

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.

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

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

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 (2008). Similarity measures for binary and numerical data : a survey. *Intern. J. of Knowledge Engineering and Soft Data Paradigms*, to appear.

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.

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

Summary

The choice of a comparison measure is a central issue for information retrieval and machine learning tasks. In this paper, we consider this problem in the case where only the order induced by the measure is of interest, and not its numerical values, as occurs e.g. for document retrieval systems. We study the classic comparison measures for numerical data, as usual distances and kernels and identify equivalent measures that always induce the same order; for non equivalent measures we quantify their disagreement through equivalence degrees based on the generalised Kendall's rank correlation, studied both from a theoretical and an experimental point of view.