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.