# 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