Consensus de classifications basé sur les regroupements fréquents

Bruno Leclerc

École des Hautes Études en Sciences Sociales
Centre d'Analyse et de Mathématique Sociales (CNRS UMR 8557)
54 boulevard Raspail, 75270 Paris cedex 06, France
leclerc@ehess.fr

Résumé. Les classifications considérées ici sont des ensembles de classes deux à deux incomparables pour l'inclusion, par exemple des partitions, ou simplement une classe unique. Soit $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$ un profil de telles classifications sur un ensemble fixé fini E, que l'on veut agréger en une classification unique $\mathcal{D}$. Pour un entier p compris entre 1 et k (inclus), on définit un *consensus par regroupements fréquents* en considérant les classes maximales incluses dans des éléments d'au moins p des $\mathcal{D}_i$. On étudie les propriétés de cette règle de consensus et on en donne trois caractérisations.

1 Introduction

Nous nous intéressons dans cet article à des regroupements d'objets apparaissant fréquemment dans une collection de classifications (produits, par exemple, par des itérations successives d'un même algorithme). On se propose d'étudier les systèmes de classes obtenus de cette façon, que nous appelons ici *regroupements fréquents*.

Des considérations provenant de divers domaines vont être évoquées. En particulier, un résultat obtenu précédemment aura un rôle important. Soient E un ensemble fini fixé et $R \subseteq (\mathcal{P}(E))^2$ une relation binaire sur l'ensemble $\mathcal{P}(E)$ des parties de E. Nous avons montré dans plusieurs travaux antérieurs (Domenach et Leclerc 2004b, Leclerc 2004, Leclerc 2005) l'unicité d'une classification $\mathcal{D}$, sous forme d'une famille de Moore (la définition d'une telle famille est donnée ci-dessous) vérifiant deux conditions relatives à R et généralisant celles posées par Adams (1986) pour le consensus d'arbres de classification ; ces conditions portent sur l'ajustement à R de la relation d'emboîtement (définie au paragraphe 5 ci-dessous) de $\mathcal{D}$. On a alors un problème d'existence, car une telle classification $\mathcal{D}$ n'existe pas pour toute relation R. En fait, Adams établit cette existence dans le cas particulier qu'il considère, celui du consensus de hiérarchies selon une forme de règle d'unanimité. Nous allons montrer que les regroupements fréquents correspondent à des fonctions de consensus du type de celle d'Adams, mais appliquées à des objets différents, et moins contraignantes que des fonctions d'unanimité. De plus, la détermination par ces fonctions d'une classification consensus $\mathcal{D}$ est proche de celle des "motifs fréquents", qui constitue un thème majeur en fouille des données pour la recherche de règles d'association (cf., e.g., Hipp et al. 2000, Han et Kamber 2001).