# Classification Conceptuelle avec Généralisation par Intervalles

Paula Brito*, Géraldine Polaillon**

*Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
mpbrito@fep.up.pt , http://www.fep.up.pt/docentes/mpbrito
**SUPELEC Science des Systèmes (E3S) - Département Informatique
Plateau de Moulon, 3 rue Joliot Curie, 91192 Gif-sur-Yvette cedex, France
geraldine.polaillon@supelec.fr

**Résumé.** Nous nous intéressons aux méthodes de classification hiérarchique ou pyramidale, où chaque classe formée correspond à un concept, i.e. une paire (extension, intension), considérant des données décrites par des variables quantitatives à valeurs réelles ou intervalles, ordinales et/ou prenant la forme de distribution de probabilités/fréquences sur un ensemble de catégories. Les concepts sont obtenus par une correspondance de Galois avec généralisation par intervalles, ce qui permet de traiter les données de différents types dans un cadre commun. Une mesure de la généralité d'un concept est alors calculée sous une forme commune pour les différents types de variables. Un exemple illustre la méthode proposée.

## 1 Introduction

La méthode de classification hiérarchique ou pyramidale proposée par Brito (voir, par exemple, Brito (1995)) permet de traiter des données multi-valuées où chaque classe formée est un concept, i.e., une paire (extension, intension). Cette méthode a été étendue par la suite à des données décrites par des variables modales, et permettant la prise en compte de l'existence de règles hiérarchiques entre variables catégoriques multi-valuées ou entre variables modales - voir Brito et De Carvalho (2008) pour une vision générale. Un critère numérique additionnel est défini, une mesure de "généralité", qui permet, à chaque étape, de choisir parmi les agrégations possibles. Les classes formées sont des "concepts", décrits en extension par la liste de ses membres, et en intension, par une expression conjonctive des valeurs prises par les variables constituant une condition nécessaire et suffisante d'appartenance à la classe. Les concepts sont obtenus grâce à la définition de correspondances de Galois pour chaque type de variables.

L'utilisation des treillis de Galois en analyse des données est d'abord due à Barbut et Monjardet (1970) et a été développée par Ganter et Wille (1999), d'abord pour des variables binaires ; cette approche a été appliquée à des variables non binaires sous condition préalable d'un recodage des données. Brito (1994) a défini des correspondances de Galois pour des variables quantitatives (réelles ou à valeurs intervalle) et qualitatives (mono ou multi-valuées), ce qui permet de traiter directement les données sans recodage ; par la suite cette approche a été étendue aux variables modales (Brito et Polaillon, 2005), i.e., variables qui prennent