Classification multi-labels graduée: Apprendre les relations entre les labels ou limiter la propagation d'erreur ?

Khalil Laghmari^{*,**}, Christophe Marsala^{**}, Mohammed Ramdani^{*}

^{*}Laboratoire Informatique de Mohammedia,
FSTM, Hassan II University of Casablanca,
BP 146 Mohammedia 20650 Maroc
laghmari.khalil@gmail.com, ramdani@fstm.ac.ma

^{**}Sorbonne Universités,
UPMC Univ Paris 06,
CNRS, LIP6 UMR 7606,
4 place Jussieu 75005 Paris, France
christophe.marsala@lip6.fr

Résumé. La classification multi-labels graduée est la tâche d'affecter à chaque donnée l'ensemble des labels qui lui correspondent selon une échelle graduelle de degrés d'appartenance. Les labels peuvent donc avoir à la fois des relations d'ordre et de co-occurrence.

D'un côté, le fait d'ignorer les relations entre les labels risque d'aboutir à des prédictions incohérentes, et d'un autre côté, le fait de prendre en compte ces relations risque de propager l'erreur de prédiction d'un label à tous les labels qui lui sont reliés.

Les approches de l'état d'art permettent soit d'ignorer les relations entre les labels, soit d'apprendre uniquement les relations correspondant à une structure de dépendance figée. L'approche que nous proposons permet l'apprentissage des relations entre les labels sans fixer une structure de dépendance au préalable. Elle est basée sur un ensemble de classifieurs mono-labels, un pour chaque label. L'idée est d'apprendre d'abord toutes les relations entre les labels y compris les relations cycliques. Ensuite les dépendances cycliques sont résolues en supprimant les relations d'intérêt minimal. Des mesures sont proposées pour évaluer l'intérêt d'apprendre chaque relation. Ces mesures permettent d'agir sur le compromis entre l'apprentissage de relations pour une prédiction cohérente et la minimisation du risque de la propagation d'erreur de prédiction.

1 Introduction

La classification multi-labels (Tsoumakas et Katakis (2007) ; Zhang et Zhou (2014)) consiste à affecter à chaque donnée un ou plusieurs labels en même temps, et la classification ordinaire (Frank et Hall (2001)) consiste à affecter à chaque donnée un label