Classification faiblement supervisée : arbre de décision probabiliste et apprentissage itératif

Riwal Lefort*,** Ronan Fablet**
Jean-Marc Boucher**

*Ifremer/STH, Technopole Brest Iroise - 29280 Plouzane, France
http://www.ifremer.fr
**Telecom Bretagne/LabSTICC, Technopol Brest Iroise
CS83818, 29238 Brest Cedex, France
riwal.lefort@telecom-bretagne.eu
http://www.telecom-bretagne.eu

Résumé. Dans le domaine de la fouille de données, il existe plusieurs types de modèles de classification qui dépendent de la complexité de l'ensemble d'apprentissage. Ce papier traite de la classification faiblement supervisée pour laquelle l'ensemble d'apprentissage est constitué de données de labels inconnus mais dont les probabilités de classification *a priori* sont connues. Premièrement, nous proposons une méthode pour apprendre des arbres de décision à l'aide des probabilités de classification *a priori*. Deuxièmement, une procédure itérative est proposée pour modifier les labels des données d'apprentissage, le but étant que les *a priori* faibles convergent vers des valeurs binaires, et donc vers un *a priori* fort. Les méthodes proposées sont évaluées sur des jeux de données issus de la base de données UCI, puis nous proposons d'appliquer ces méthodes d'apprentissage dans le cadre de l'acoustique halieutique.

1 Introduction

Dans le domaine de la fouille de données, de nombreuses applications nécessitent le développement de modèles de classification stables et robustes. On peut citer par exemple, la reconnaissance d'objets (Crandall and Huttenlocher, 2006), la reconnaissance de texture d'images (Lazebnik et al., 2005), la reconnaissance d'évènement dans des vidéos (Hongeng et al., 2004), ou encore l'analyse de scènes dans des images (Torralba, 2003). Ce type de problème correspond à deux étapes principales : la définition d'un vecteur de descripteurs qui décrit l'objet considéré et le développement d'un modèle de classification pour les objets considérés. La seconde étape requière une phase d'apprentissage dont le procédé dépend des caractéristiques de l'ensemble d'apprentissage. Par exemple, en classification semi-supervisée (Chapelle et al., 2006), l'ensemble d'apprentissage est constitué de quelques données labélisées, complétées par un ensemble conséquent d'exemples sans label. Pour certaines applications, les performances de classification sont alors identiques au cas de l'apprentissage supervisé.

De manière plus générale, ce papier traite de l'apprentissage faiblement supervisé qui inclue à la fois l'apprentissage supervisé, l'apprentissage semi-supervisé, et l'apprentissage non