SLEMC : Apprentissage semi-supervisé enrichi par de multiples clusterings.

Cédric Wemmert*, Germain Forestier*

*Université de Strasbourg
LSIIT UMR 7005 CNRS/UDS
{wemmert,forestier}@unistra.fr

Résumé. La tâche de classification supervisée consiste à induire un modèle de prédiction en utilisant un ensemble d'échantillons étiquetés. La précision du modèle augmente généralement avec le nombre d'échantillons disponibles. Au contraire, lorsque seuls quelques échantillons sont disponibles pour l'apprentissage, le modèle qui en résulte donne généralement des résultats médiocres. Malheureusement, comme la tâche d'étiquetage est souvent très coûteuse en temps, les utilisateurs fournissent principalement un très petit nombre d'objets étiquetés. Dans ce cas, de nombreux échantillons non étiquetés sont généralement disponibles. Lorsque les deux types de données sont disponibles, les méthodes d'apprentissage semi-supervisé peuvent être utilisées pour tirer parti à la fois des données étiquetées et non étiquetés. Dans cet article nous nous concentrons sur le cas où le nombre d'échantillons étiquetés est très limité. Ainsi, définir des approches capables de gérer ce type de problème représente un verrou scientifique important à lever. Dans cet article, nous passons en revue et comparons les différents algorithmes d'apprentissage semi-supervisé existant, et nous présentons également une nouvelle manière de combiner apprentissage supervisé et non supervisé afin d'utiliser conjointement les données étiquetées et non étiquetées. La méthode proposée utilise des résultats de clustering pour produire des descripteurs supplémentaires des données, qui sont utilisés alors lors d'une étape d'apprentissage supervisé. L'efficacité de la méthode est évaluée sur différents jeux de données de l'UCI [1] et sur une application réelle de classification d'une image de télédétection à très haute résolution avec un nombre très faible d'échantillons étiquetés. Les expériences donnent des indications sur l'intérêt de combiner des données étiquetées et non dans le cadre de l'apprentissage automatique.

1 Introduction

Le nombre d'exemples étiquetés est un paramètre essentiel lors de la phase d'apprentissage en classification supervisée. Si trop peu d'exemples sont disponibles, le modèle prédictif induit par l'apprentissage aura des performances relativement faibles. Malheureusement, dans

1. Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.