Détection non supervisée d'une sous-population par méthode d'ensemble et changement de représentation itératif

Christine Martin, Antoine Cornuéjols

AgroParisTech, département MMIP et INRA UMR-518
16, rue Claude Bernard
F-75231 Paris Cedex 5 (France)
christine.martin , antoine.cornuejols@agroparistech.fr,
http://www.agroparistech.fr/mia/equipes:membres:page:christine

Résumé. L'apprentissage non supervisé a classiquement pour objectif la détection de sous-populations homogènes (classes) considérées de manière équivalente sans information *a priori* sur celles-ci. Le problème étudié dans cet article est quelque peu distinct. On se focalise ici uniquement sur une sous-population d'intérêt que l'on cherche à identifier avec un rappel et une précision optimales.

Nous proposons, pour cela, une méthode s'appuyant sur les principes suivants : (1) travailler dans l'espace de représentation fourni par des experts faibles pour cette tâche, (2) confronter ces experts pour détecter des seuils de sélection plus pertinents, et (3) les combiner itérativement afin de converger vers l'expert idéal. Cette méthode est éprouvée et comparée sur des données synthétiques.

1 Introduction

De nombreuses tâches de fouille de données ou d'apprentissage impliquent la détection d'un sous-ensemble inconnu d'éléments dans une collection d'éléments non étiquetés.

Classiquement des méthodes de regroupement ou *clustering* sont utilisées dans l'espoir de faire apparaître la classe d'intérêt. Cependant les résultats sont généralement très sensibles à la technique utilisée et aux paramètres choisis. Une autre approche est d'utiliser une *méthode de filtre* (Sahai (2000) ; Kira et Rendell (1992)) afin d'évaluer la qualité des objets en fonction d'un critère supposé pertinent et d'ordonner ainsi les objets en qualité décroissante puis de déterminer un seuil dans ce classement permettant de distinguer les deux classes. Malheureusement, ces méthodes donnent des résultats très variables en fonction, d'une part, de leur adéquation avec les régularités cibles inconnues, et, d'autre part, des seuils de sélection choisis.

Nous proposons de circonvenir ces difficultés en utilisant une méthode d'ensemble basée sur des méthodes de filtres nommées ici *« experts faibles »*. A l'image des méthodes d'ensemble proposées en apprentissage supervisé, telles que le boosting (Freund (1995)), nous combinons des méthodes de filtres, pour converger vers la méthode idéale, c'est-à-dire permettant une détection aisée des objets considérés.

Le problème étudié ici s'apparente donc à celui du tri (*ranking* en anglais) d'un ensemble d'objets. Pour cela, des approches par apprentissage d'une fonction de score évaluant chaque