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.

Summary

Unsupervised learning seeks the detection of somewhat homogeneous sub-populations of a given set when no label is available. Each class or group is considered as equivalent. In this paper, we tackle a different, if related, problem. Based solely on the assumption that there exists a sub-population that interests us in the set of instances, we want to identify its member as well as possible.

The method presented here is based on three ideas: (1) using a change of representation whereby the objects are represented in the space of the evaluations of "weak experts", (2) through the study of experts in pairs, to apply a non linear transformation of the evaluations of each expert in order to amplify their tendency to discriminate objects of interest from the other, and (3) to combine iteratively the available "weak experts" to get a better final combined expert. The method has been tested and compared on synthetic data showing improvements in all cases.