Sous-échantillonnage topographique par apprentissage semi-supervisé

Mustapha Lebbah, Younès Bennani

LIPN UMR CNRS 7030, Université Paris 13,
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse
prénom.nom@lipn.univ-paris13.fr

Résumé. Plusieurs aspects pourraient influencer les systèmes d'apprentissage existants. Un de ces aspects est lié au déséquilibre des classes dans lequel le nombre d'observations appartenant à une classe, dépasse fortement celui des observations dans les autres classes. Dans ce type de cas assez fréquent, le système d'apprentissage a des difficultés au cours de la phase d'entraînement liées au déséquilibre inter-classe. Nous proposons une méthode de sous-échantillonnage adaptatif pour traiter ce type de bases déséquilibrées. Le processus procède par le sous-échantillonnage des données majoritaires, guidé par les données minoritaires tout au long de la phase d'un apprentissage semi-supervisée. Nous utilisons comme modèle d'apprentissage les cartes auto-organisatrices. L'approche proposée a été validée sur plusieurs bases de données en utilisant les arbres de décision comme classificateur avec une validation croisée. Les résultats expérimentaux ont montré des performances très prometteuses.

1 Introduction

La plupart des algorithmes d'apprentissage sont basés sur deux hypothèses : La première est le critère à minimiser qui est le nombre d'erreurs. La deuxième est que les données d'apprentissage doivent êtres un échantillon représentatif de la population sur laquelle le modèle sera appliqué. Ces deux hypothèses ne sont pas respectées pour certains modèles quand ils sont construits à partir de données déséquilibrées. Nous pouvons l'illustrer par un exemple simple pris souvent en littérature : si 99% des données appartiennent à une seule classe, il sera difficile de faire mieux que le 1% d'erreur obtenue en classant tous les individus dans cette classe. Il convient donc de trouver d'autres solutions et hypothèses adaptées au problème de déséquilibre sans remettre en cause les fondements des algorithmes. Weiss (2003) propose de distinguer six catégories de problèmes liés aux données déséquilibrées, et à l'apprentissage des classes rares. Ces catégories sont (Marcellin et al. (2008)) : (**a**) Métriques inappropriées : dans ce cas, les mesures utilisées au cours du processus d'apprentissage ne sont pas adaptées aux classes déséquilibrées. (**b**) Manque "absolu" de données : ce problème est observé lorsque les données disponibles ne sont pas assez suffisantes pour définir clairement les frontières de la classe. (**c**) Manque "relatif" de données : c'est un problème similaire au manque absolu, sauf que dans ce cas ce manque est relatif à la taille de la base de données majoritaires. (**d**) Fragmentation des données : ce problème est lié aux algorithmes ayant une approche descendante, qui partent de l'espace de tous les individus et le partitionnent récursivement en sous-espaces