

Sélection et transformation de variables pour la classification Multi-Label par une approche MDL

Sènam C. Fréjus Ahomagnon *, Nicolas Voisine**
Marc Boullé**

*Polytech Nantes
jusberlin@gmail.com
**Orange Labs Lannion
prenom.nom@orange.com

Résumé. La classification multi-label est une extension de la classification supervisée au cas de plusieurs labels. Elle a connu un regain d'intérêt récent dans la communauté du machine learning de par son utilité dans plusieurs domaines. Comme pour tout problème de machine learning, le besoin de prétraiter les données multi-label est apparu comme une nécessité afin d'améliorer les performances des classifieurs. Dans cet article, nous introduisons une nouvelle méthode permettant de prétraiter des variables descriptives par discrétisation ou groupement de valeur, dans le cas de plusieurs labels à prédire. Le choix du meilleur prétraitement est posé comme un problème de sélection de modèle, et est résolu au moyen d'une approche bayésienne. Une étude comparative est réalisée avec d'autres méthodes de l'état de l'art afin de positionner la nouvelle méthode et de montrer l'intérêt de la sélection de variables pour la classification.

1 Introduction

Cet article se place dans le cadre de la classification multi-label où l'on veut prédire un ensemble de labels pour une instance. La classification de sons, de vidéos et de textes fait partie de ces problèmes où un élément peut être associé à plusieurs labels. En prenant par exemple un article sur le Dalai Lama : en classification de texte, il peut être associé à la politique et à la religion. Suite à l'accroissement des gisements de données multi-label, le problème de la grande dimensionnalité dans les données s'est posé. Comme dans tout problème de machine learning, la mise en place de méthodes de prétraitement des données est donc apparue comme une nécessité. Des études variées [Dendamrongvit et al. (2011), Trohidis et al. (2008), Spolaôr et al. (2012), Zhao et al. (2010)] ont montré que la réduction de l'espace des variables explicatives pouvait être faite sans détériorer les performances des classifieurs.

Différentes études ayant abondé dans ce sens ont abouti à des méthodes de sélection de variables pour les classifieurs multi-label. Par analogie à la typologie des méthodes de classification multi-label (Madjarov et al., 2012), trois grandes familles de méthodes de sélection peuvent être dégagées de ces études. La première famille est celle des méthodes de sélection de variables par transformation du problème, où en transformant le problème de sélection multi-label en plusieurs problèmes de sélection mono-label (BR) ou multi-classe (LP), les résultats