Mesure non symétrique pour l'évaluation de modèles, utilisation pour les jeux de données déséquilibrés

Julien Thomas[*,**], Pierre-Emmanuel Jouve[**], Nicolas Nicoloyannis[*]

[*]Laboratoire ERIC, Université Lumière Lyon2, France
http://eric.univ-lyon2.fr
[**]Société Fenics Lyon, France
http://www.fenics-sas.com

Résumé. Les critères servant à l'évaluation de modèles d'apprentissage supervisé ainsi que ceux utilisés pour bâtir des arbres de décision sont, pour la plupart, symétriques. De manière pragmatique, cela signifie que chacune des modalités de la variable endogène se voit assigner une importance identique. Or, dans nombre de cas pratiques cela n'est pas le cas. Ainsi, on peut notamment prendre l'exemple de jeux de données fortement déséquilibrés pour lesquels l'objectif principal est l'identification des objets représentatifs de la modalité minoritaire (Aide au diagnostic, identification de phénomènes inhabituels : fraudes, pannes...). Dans ce type de situation il apparaît clairement qu'assigner une importance identique aux erreurs de prédiction ne constitue pas la meilleure des solutions. Nous proposons dans cet article un critère (pouvant servir à la fois pour l'évaluation de modèles d'apprentissage supervisé ou encore de critère utilisé pour bâtir des arbres de décision) prenant en compte cet aspect non symétrique de l'importance associée à chacune des modalités de la variable endogène. Nous proposons ensuite une évolution des modèles de type forêts aléatoires utilisant ce critère pour les jeux de données fortement déséquilibrés.

1 Introduction

L'évaluation des performances d'un modèle constitue l'étape finale de tout processus d'apprentissage supervisé. Elle est le retour nécessaire à l'utilisateur pour le guider dans la poursuite de sa fouille de données. Ces mesures, comme celles utilisées pour bâtir des arbres de décisions, sont généralement symétriques. De façon pratique, on entend par symétrique le fait que les erreurs sur chaque modalité de la variable endogène se voient attribuer une importance similaire. Or de nombreux exemples industriels nous montrent que cela n'est pas toujours le cas, en particulier lorsqu'on se trouve en présence de jeux de données fortement déséquilibrés : aide au diagnostic (Grzymala-Busse, 2000), identification de phénomènes inhabituels comme les fraudes lors des transactions par cartes bancaires (Chan, 2001) ou les pannes d'équipements de télécommunications (Weiss, 1998), et bien d'autres encore. Dans ce type de cas l'objectif principal est d'identifier les instances représentatifs de la classe minoritaire.
Il est pour cela nécessaire d'utiliser des méthodologies d'apprentissage adaptées (Weiss, 2004)