Améliorer les performances d'un modèle prédictif: perspectives et réalité

Stéphane TUFFERY

6, rue Gaston Turpin - 44000 Nantes
stephane.tuffery@univ-rennes1.fr, data.mining@free.fr
http://data.mining.free.fr

Résumé. Dans cet article, nous montrons que les performances d'un modèle prédictif dépendent généralement plus de la qualité des données et du soin apporté à leur préparation et à leur sélection, que de la technique de modélisation elle-même. Entre deux techniques, l'écart de performance est souvent négligeable en regard des incertitudes résultant de la définition de la variable à expliquer et de la représentativité de l'échantillon d'étude. Toutefois, le rééchantillonnage et l'agrégation de modèles peuvent permettre de réduire drastiquement la variance et parfois même le biais de certains modèles. De bons résultats peuvent aussi être obtenus simplement par la partition de modèles, c'est-à-dire en partitionnant en classes l'échantillon initial et en construisant un modèle sur chaque classe.

1 Introduction

Le foisonnement de découvertes statistiques de ces dernières années (modèle additif généralisé, séparateurs à vaste marge, régression logistique PLS, etc.) ne doit pas nous faire oublier que dans la plupart des problèmes réels de classement[1], notamment ceux qui se posent dans l'assurance, la banque et le marketing[2], les performances d'un modèle dépendent souvent moins de la technique de modélisation que de la nature et de la qualité des données. De nombreux comparatifs basés sur un ou plusieurs jeux de données mettent en évidence des écarts de performance très limités entre les techniques. Ces écarts sont si ténus qu'ils sont parfois dérisoires en regard des incertitudes qui pèsent sur eux. Première incertitude : la définition de la variable à expliquer, qui est parfois beaucoup moins naturelle qu'elle peut l'être dans des domaines comme la médecine. Deuxième incertitude : la représentativité de l'échantillon d'étude (dont sont extraits les échantillons d'apprentissage et de test) par rapport à une population dont tous les individus n'ont pas été observés (biais de sélection) ou qui a pu évoluer depuis l'observation (la modélisation s'appuie sur des échantillons rétrospectifs). Troisième incertitude :

[1] Attention à la terminologie : les statisticiens francophones appellent "classement" (technique supervisée) ce que les anglo-saxons- et certains data miners français- appellent "classification". Quand au terme français "classification" (technique non supervisée), il se traduit en anglais par "clustering".

[2] Nous parlons de ces domaines qui sont le sujet du présent numéro de la RNTI et qui sont aussi du ressort de l'auteur ; certaines conclusions sont généralisables à d'autres domaines, mais pas toutes : ainsi la parcimonie dans le nombre de variables du modèle est moins pertinente dans certains domaines comme la génomique ou la chimiométrie.