Sélection supervisée d'instances : une approche descriptive

Sylvain Ferrandiz[*,**], Marc Boullé[*]

[*]France Télécom R&D,
2, avenue Pierre Marzin, 22300 Lannion
sylvain.ferrandiz@francetelecom.com,
marc.boulle@francetelecom.com,
[**]GREYC, Université de Caen,
boulevard du Maréchal Juin, BP 5186, 14032 Caen Cedex,

Résumé. La classification suivant le plus proche voisin est une règle simple et performante. Sa mise en oeuvre pratique nécessite, tant pour des raisons de coût de calcul que de robustesse, de sélectionner les instances à conserver. La partition de Voronoi induite par les prototypes constitue la structure sous-jacente à cette règle. Dans cet article, on introduit un critère descriptif d'évaluation d'une telle partition, quantifiant le compromis entre nombre de cellules et discrimination de la variable cible entre les cellules. Une heuristique d'optimisation est proposée, tirant partie des propriétés des partitions de Voronoi et du critère. La méthode obtenue est comparée avec les standards sur une vingtaine de jeux de données de l'UCI. Notre technique ne souffre d'aucun défaut de performance prédictive, tout en sélectionnant un minimum d'instances. De plus, elle ne sur-apprend pas.

1 Introduction

La classification supervisée constitue un problème d'apprentissage classique. On dispose dans ce cas, en plus des variables descriptives (ou endogènes), d'une variable cible (ou exogène). En phase d'exploration des données, c'est la dépendance de la variable cible vis-à-vis des variables descriptives qu'on vise à expliciter. En phase de modélisation, le but est de fournir la meilleure prédiction possible pour toute nouvelle instance à classifier. Quelle que soit la situation, la connaissance est à extraire d'un échantillon de N instances étiquetées.

Une méthode de classification usuelle est la règle de classification suivant le plus proche voisin introduite par Fix et Hodges (1951). Elle consiste à attribuer à une instance l'étiquette de l'instance la plus proche parmi celles constituant l'échantillon. La mise en œuvre de cette modélisation soulève deux questions fondamentales :

– Quelle mesure de similitude employer ?
– Quelles instances de l'échantillon conserver ?

La première question couvre plusieurs champs d'investigation : gestion de la présence jointe de variables continues et symboliques, normalisation des variables continues, prétraitement des variables symboliques, pondération de la contribution des variables, etc. Dans le cas continu, l'usage a consacré l'emploi de la distance euclidienne et des distances L_p ($p \geq 1$)