Sélection topologique de variables dans un contexte de discrimination

Fatima-Zahra Aazi*, Rafik Abdesselam**

*Laboratoires ERIC & LAMSAD
Universités Lumière Lyon 2, France & Hassan 1^{er}, Settat, Maroc
5, avenue Pierre Mendès-France, 69676 Bron Cedex, France
Fatima-Zahra.Aazi@univ-lyon2.fr
**COACTIS-ISH, Université de Lyon, Lumière Lyon 2
14/16, avenue Berthelot, 69363 Lyon Cedex 07, France
rafik.abdesselam@univ-lyon2.fr
http://eric.univ-lyon2.fr/~rabdesselam/fr/

Résumé. En apprentissage automatique, la présence d'un grand nombre de variables explicatives conduit à une plus grande complexité des algorithmes et à une forte dégradation des performances des modèles de prédiction. Pour cela, une sélection d'un sous-ensemble optimal discriminant de ces variables s'avère nécessaire. Dans cet article, une approche topologique est proposée pour la sélection de ce sous-ensemble optimal. Elle utilise la notion de graphe de voisinage pour classer les variables par ordre de pertinence, ensuite, une méthode pas à pas de type ascendante "forward" est appliquée pour construire une suite de modèles dont le meilleur sous-ensemble est choisi selon son degré d'équivalence topologique de discrimination. Pour chaque sous-ensemble, le degré d'équivalence est mesuré en comparant la matrice d'adjacence induite par la mesure de proximité choisie à celle induite par la "meilleure" mesure de proximité discriminante dite de référence. Les performances de cette approche sont évaluées à l'aide de données simulées et réelles. Des comparaisons de sélection de variables en discrimination avec une approche métrique montrent une bien meilleure sélection à partir de l'approche topologique proposée.

1 Introduction

Le changement majeur de la nature des données, notamment l'accroissement de leur masse et de leur taille, nécessite de nouvelles approches de traitement statistique adaptées aux caractéristiques de ces données modernes et massives. En particulier, la grande dimension des données (nombre important de variables) pose un certain nombre de problèmes à la statistique multivariée, notamment aux techniques prédictives de classement. On peut citer les problèmes numériques, de collinéarité, d'inférence ou de biais des estimateurs. Il est nécessaire de développer des méthodes capables de pallier ces problèmes.

On s'intéresse ici au problème de sélection de variables en discrimination. En effet, vu que les variables pertinentes ne sont pas connues *a priori*, la sélection est justifiée, en présence