Khiops: outil d'apprentissage supervisé automatique pour la fouille de grandes bases de données multi-tables

Marc Boullé*

*2 avenue Pierre Marzin,
22300 Lannion, France,
marc.boulle@orange.com,
http://www.marc-boulle.fr

Résumé. Khiops est un outil d'apprentissage supervisé automatique pour la fouille de grandes bases de données multi-tables. L'importance prédictive des variables est évaluée au moyen de modèles de discrétisation dans le cas numérique et de groupement de valeurs dans le cas catégoriel. Dans le cas d'une base multi-tables, par exemple des clients avec leurs achats, une table d'analyse individus $\times$ variables est produite par construction automatique de variables. Le modèle de classification utilisé est un classifieur Bayésien naïf avec sélection de variables et moyennage de modèles. L'outil est adapté à l'analyse des grandes bases de données, avec des millions d'individus, des dizaines de milliers de variables et des centaines de millions d'enregistrements dans les tables secondaires.

1 Introduction

Dans un projet de fouille de données, la phase de préparation des données vise à extraire une table de données pour la phase de modélisation (Pyle, 1999). La préparation des données est non seulement coûteuse en temps d'étude, mais également critique pour la qualité des résultats escomptés. Dans le cas de la fouille de données à Orange, le contexte industriel impose des contraintes telles que le potentiel des données collectées est largement sous-utilisé.

La préparation repose essentiellement sur la recherche d'une représentation pertinente pour le problème à modéliser, recherche qui se base sur des étapes complémentaires de construction et de sélection de variables. La sélection de variables a été largement étudiée dans la littérature (Guyon et al., 2006). La construction de variables (Liu et Motoda, 1998) est un sujet nettement moins étudié, qui représente néanmoins un travail considérable pour l'analyste de données. Celui-ci exploite sa connaissance du domaine pour créer de nouvelles variables potentiellement informatives. En pratique, les données initiales sont souvent issues de bases de données relationnelles et ne sont pas directement exploitables pour la plupart des techniques de classification qui exploitent un format tabulaire, avec en lignes les individus à analyser et en colonnes les variable. Par exemple, dans le domaine de la gestion de la relation client pour des problèmes de type prédiction d'attrition (passage à la concurrence) ou d'appétence à un produit ou service, les données disponibles par client sont multiples et volumineuses : age, genre, adresse, données INSEE, détails de communication, logs d'usage des produits et services...