Khiops: outil de préparation et modélisation des données pour la fouille des grandes bases de données

Marc Boullé*

*2 avenue Pierre Marzin
marc.boulle@orange-ftgroup.com,
http://perso.rd.francetelecom.fr/boulle/

Résumé. Khiops est un outil de préparation des données et de modélisation pour l'apprentissage supervisé et non supervisé. L'outil permet d'évaluer de façon non paramétrique la corrélation entre tous types de variables dans le cas non supervisé et l'importance prédictive des variables et paires de variables dans le cas de la classification supervisée. Ces évaluations sont effectuées au moyen de modèles de discrétisation dans le cas numérique et de groupement de valeurs dans le cas catégoriel, ce qui permet de rechercher une représentation des données efficace au moyen d'un recodage des variables. L'outil produit également un modèle de scoring pour les tâches d'apprentissage supervisé, selon un classifieur Bayesien naif avec sélection de variables et moyennage de modèles.

L'outil est adapté à l'analyse des grandes bases de données, avec des centaines de milliers d'individus et des dizaines de milliers de variables, et a permis de participer avec succès à plusieurs challenges internationaux récents.

Présentation de l'outil

La phase de préparation des données est particulièrement importante dans le processus de fouille de données (Pyle, 1999). Elle est critique pour la qualité des résultats, et consomme typiquement de l'ordre de 80% du temps d'une étude de fouille de données. Dans le cas de la fouille de données à France Télécom, le contexte industriel impose des contraintes telles que le potentiel des données collectées dans les systèmes d'information est largement sous-utilisé.

L'outil Khiops intègre les travaux sur les modèles en grille (Boullé, 2006, 2007a,b) et les diffuse dès qu'ils ont atteint une maturité suffisante. Dans le cas univarié, un modèle en grille s'apparente à une discrétisation pour une variable numérique et à un groupement de valeurs pour une variable catégorielle. Dans le cas multivarié, chaque variable est partitionnée en intervalles ou groupes de valeurs selon sa nature numérique ou catégorielle. L'espace complet des données est alors partitionné en une grille de cellules résultant du produit cartésien des partition univariées. Ces modèles permettent alors une estimation non paramétrique de densité conditionnelle dans le cas supervisé et jointe dans le cas non supervisé. La granularité optimale de la grille des données est recherchée en appliquant une approche Bayesienne de la sélection de modèles et en exploitant des algorithmes sophistiqués d'optimisation combinatoire.