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.

L'outil Khiops est centré sur la préparation des données et la modélisation. Il ne s'agit pas d'un atelier de data mining : on n'y trouve ni connecteurs spécialisés vers des bases de données, ni interface de visualisation avancée des résultats d'analyse.

La version actuellement diffusée comprend les fonctionnalités principales suivantes :

- préparation des données en supervisé par discrétisation et groupement de valeurs,
- préparation des données en non supervisé par évaluation non paramétrique de la corrélation des paires de variables au moyen de modèles en grilles bivariés,
- modélisation par classifieur Bayesien naïf, avec prétraitements univariés et bivariés, sélection de variables et moyennage de modèles.

L'outil est écrit en langage C++ pour la partie algorithmique, et en Java pour l'interface graphique. Il est utilisable à la fois en mode interface graphique et en mode batch, ce qui permet de l'intégrer aisément dans une chaîne de traitements.

La version actuellement diffusée est utilisée en interne à France Telecom dans de nombreux domaines applicatifs : marketing client (modèles de churn, d'appétence au nouveaux services...), text mining, web mining, réseaux sociaux, étude technico-économique, caractérisation du trafic internet, ergonomie, sociologie des usages....

Cette version est également diffusée en externe sous forme de shareware, sur le site `http ://www.francetelecom.com/en/group/rd/offer/software/applications/providers/khiops.html`.

Références

Boullé, M. (2006). MODL : a Bayes optimal discretization method for continuous attributes. *Machine Learning 65*(1), 131–165.

Boullé, M. (2007a). Compression-based averaging of selective naive Bayes classifiers. *Journal of Machine Learning Research 8*, 1659–1685.

Boullé, M. (2007b). *Recherche d'une représentation des données efficace pour la fouille des grandes bases de données*. Ph. D. thesis, ENST.

Pyle, D. (1999). *Data preparation for data mining*. Morgan Kaufmann Publishers, Inc. San Francisco, USA.

Summary

Khiops is a data preparation and modeling tool for supervised and unsupervised learning. It exploits non parametric models to evaluate the correlation between any type of variables in the unsupervised case and the predictive importance of input variables and pairs of input variables in the supervised case. These evaluations are performed by the mean of discretization models in the numerical case and of value grouping models in the categorical case, which correspond to the search for an efficient data representation owing to variable recoding. The tool also produces a scoring model for supervised learning tasks, according to a naive Bayes approach, with variable selection and model averaging.

The tools is designed for the management of large datasets, with hundreds of thousands of instances and tens of thousands of variables, and was succesfully evaluated in several international data mining challenges.