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...

Dans ce contexte, la construction de variables agrégeant les données de log est une étape nécessaire pour produire une table d'analyse individus $\times$ variables résumant les données, ce qui représente un processus extrêmement long et complexe à mettre en oeuvre.

La fouille de données relationnelles, en anglais Multi-Relational Data Mining (MRDM), introduite par (Knobbe et al., 1999), vise à exploiter directement le formalisme multi-tables, en transformant la représentation relationnelle. En programmation logique inductive (ILP) (Džeroski et Lavrač, 2001), les données sont recodées sous forme de prédicats logiques. D'autres méthodes, dénommées propositionalisation (Kramer et al., 2001), effectuent une mise à plat au format tabulaire de la représentation multi-tables par création de nouvelles variables. Par exemple, la méthode Relaggs (Krogel et Wrobel, 2001) exploite des fonctions de type moyenne, médiane, min, max pour résumer les variables numériques des tables secondaires ou des effectifs par valeur pour les variables catégorielles secondaires. La méthode Tilde (Blockeel et al., 1998), permet de construire des agrégats complexes exploitant des conjonctions de conditions de sélection d'individus dans les tables secondaires. L'expressivité de ces méthodes se heurte néanmoins aux problèmes suivants : complexité du paramétrage de la méthode, explosion combinatoire du nombre de variables construites difficile à maitriser et risque de surapprentissage croissant avec le nombre de variables produites.

L'objectif de l'outil Khiops est d'automatiser de façon simple, efficace et robuste la construction et sélection des variables ainsi que la modélisation pour l'apprentissage supervisé dans le cas de bases de données multi-tables de grande taille.

2 Présentation de l'outil

L'outil Khiops intègre les travaux effectué à Orange Labs sur la préparation des données, la construction automatique de variables pour les bases multi-tables et la modélisation en grande volumétrie. La préparation des données se fait au moyen d'une discrétisation supervisée (Boullé, 2006) pour les variables numériques et d'un groupement de valeurs supervisé (Boullé, 2005) pour les variables catégorielles. Les méthodes associées exploitent une approche Bayésienne de sélection de modèle pour construire le modèle de préparation le plus probable connaissant les données, ce qui permet d'obtenir une estimation précise et robuste de la densité conditionnelle univariée par variable descriptive. La construction automatique de variables dans le cas multi-tables constitue l'apport majeur de l'outil. Elle se base sur la description d'un schéma en étoile [1] , avec une table racine contenant les individus à analyser (par exemple des clients) et des tables secondaires en relation 0-1 ou 0-n contenant des enregistrements complétant la description des individus (par exemple, des détails de communication). Le seul paramètre utilisateur est alors le nombre de variables à construire, par application systématique de fonctions de sélection ou d'agrégation. La méthode utilisée (Boullé, 2014) exploite une approche de régularisation Bayésienne sur la base d'une distribution a priori parcimonieuse sur l'ensemble potentiellement infini de toutes les variables pouvant être construites. Les variables sont alors construites grâce à un algorithme d'échantillonnage efficace selon cette distribution a priori. La méthode résultante est simple à utiliser, efficace en temps de calcul et robuste au

1. La terminologie utilisée est proche de celle des entrepôts de données : schéma en étoile avec table de faits et tables dimensions. Cependant, il ne s'agit pas ici de concepts de structuration d'un entrepôt de données, mais de description des individus d'une analyse statistique, avec leurs variables simples provenant de la table racine et multiples, provenant des tables secondaires et disponibles sous forme de séries de valeurs de longueur variable.

problème du sur-apprentissage. La modélisation exploite l'ensemble des variable initiales ou construites après leur préparation et les combine au moyen d'un classifieur Bayésien naïf avec sélection de variables et moyennage de modèles (Boullé, 2007, 2009).

La version Khiops V8 diffusée comprend les fonctionnalités principales suivantes :
- prise en compte des schémas multi-tables en étoile,
- construction automatique de variables pour créer une table individus $\times$variables,
- préparation des données supervisée par discrétisation et groupement de valeurs,
- modélisation par classifieur Bayésien naïf, avec prétraitements univariés, sélection de variables et moyennage de modèles,
- déploiement des modèles directement sur des base multi-tables.

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. Un outil de visualisation interactive est également disponible pour inspecter les résultats de préparation, modélisation et évaluation.

La version diffusée est utilisée en interne à Orange dans de nombreux domaines applicatifs : marketing client (modèles d'attrition, d'appétence aux nouveaux services...), fouille de texte, fouille du web, réseaux sociaux, études technico-économiques, caractérisation du trafic internet, ergonomie, sociologie des usages... Elle a été utilisée avec des bases d'apprentissage comportant des millions d'individus et des centaines de millions d'enregistrements secondaires Cette version est également téléchargeable en externe sur le site `http://www.khiops.com` avec une période d'évaluation gratuite. L'outil est disponible sous Windows et Linux, 32 et 64 bits, avec un installeur, une documentation complète, un tutoriel et des bases d'exemple.

3 Exemple d'utilisation

L'utilisation de l'outil dans le cas d'une base multi-tables est illustré ici sur un exemple simple fourni avec l'outil Khiops.

3.1 Problème à analyser

Le jeu de données utilisé *Splice Junction* (Asuncion et Newman, 2007) correspond à un problème de biologie moléculaire, plus précisément de reconnaissance de jonction entre des portions de séquences ADN de type intron ou exon dans les gènes. La classe cible contient trois valeurs : *IE* pour une jonction entre intron et exon, *EI* entre exon et intron et *N* en l'absence de jonction. Chaque instance est représentée par une séquence ADN de taille 60 centrée autour du point de jonction potentiel.

3.2 Format des données

Le jeu de donnée *Splice Junction* se représente naturellement au format multi-tables. La table principale comporte 3190 instances avec deux variables : *SampleId* pour l'identifiant de la séquence ADN et *Class* pour la variable à prédire. La table secondaire, en lien 0-n avec la table principale, comporte 191400 enregistrements, avec trois variables : *SampleId* pour faire la jointure avec la table principale, *Pos* et *Char* pour représenter la séquence ADN.

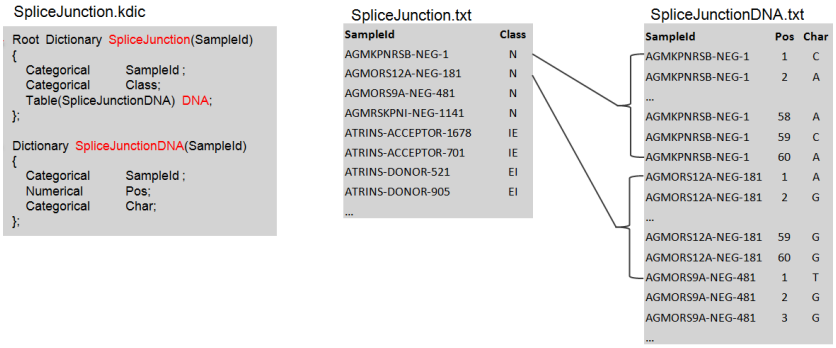

FIG. 1 – *Fichier dictionnaire et fichiers de données.*

Le schéma multi-tables est spécifié au moyen d'un fichier dictionnaire, qui décrit la composition de chaque table ainsi que la structure du schéma en étoile. Chaque table est un fichier texte, avec une ligne par enregistrement, une ligne d'en-tête optionnelle et un séparateur de champ (tabulation par défaut). La figure 1 présente le fichier dictionnaire et les fichiers de données pour le jeu de données *Splice Junction*.

3.3 Apprentissage

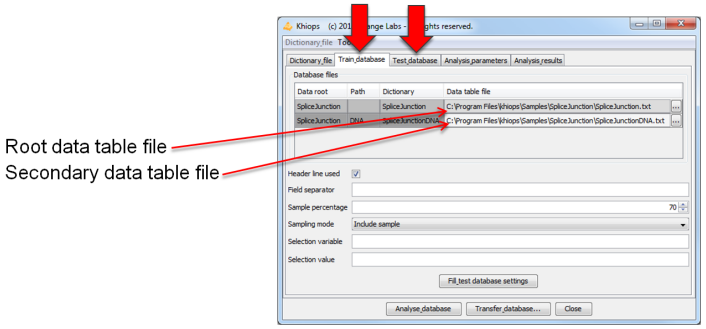

FIG. 2 – *Spécification des bases d'apprentissage et test.*

L'outil étant automatique, il est utilisable avec un minimum de paramétrage. Pour le problème *Splice Junction*, il suffit de spécifier principalement les objectifs de l'analyse :

- dictionnaire décrivant le schéma de données,
- fichiers de données en apprentissage et en test,
- variable cible,
- nombre de variables à construire (seul véritable paramètre utilisateur).

La figure 2 présente l'interface utilisateur de l'outil Khiops lors de la phase de spécification des fichiers de la base multi-tables en apprentissage et test.

3.4 Visualisation des résultats

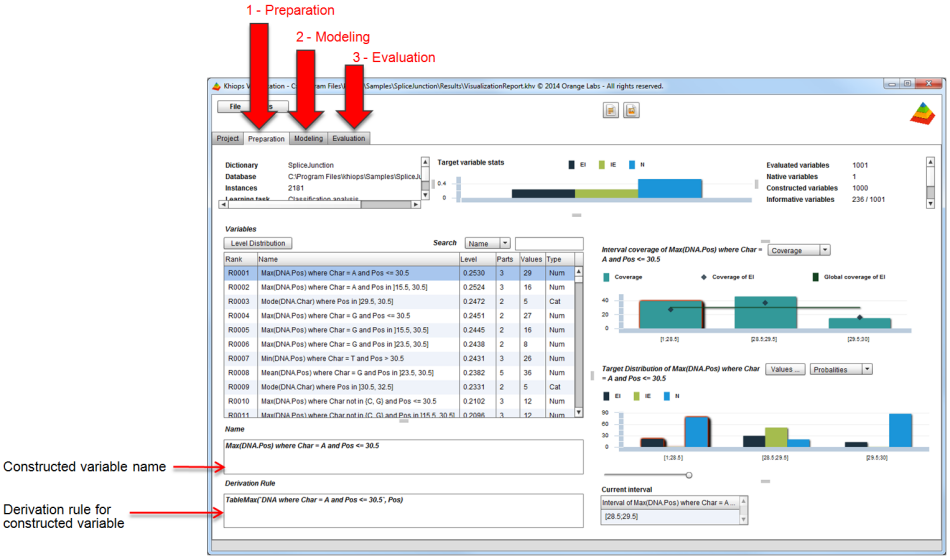

FIG. 3 – *Visualisation des résultats de préparation des données.*

On a demandé ici la construction de 1000 variables et obtenu, automatiquement et sans connaissance du domaine, un taux de bonne prédiction en test de 96%, au niveau de l'état de l'art. L'outil de visualisation permet d'inspecter les résultats d'analyse : préparation, modélisation et évaluation. Par exemple, la figure 3 présente l'interface de cet outil pour la préparation des données. Ici la variable la plus informative qui est construite a pour nom *Max(DNA.Pos) where Char = A and Pos <= 30.5*. Elle peut s'interpréter simplement comme la dernière position d'un nucléotide de type A (adénine) dans la première moitié de la séquence ADN. La discrétisation associée, en trois intervalles, montre que lorsque cette position est 29 (intervalle central), il y a un mélange des trois classes à prédire avec une sur-représentation de la classe *IE* (en vert), et sinon, la classe *IE* est absente et la classe *N* (en bleu) est largement majoritaire.

4 Conclusion

L'objectif principal de l'outil Khiops est l'automatisation du processus d'analyse de données, de façon à permettre de répondre simplement, rapidement et efficacement aux besoins croissants d'analyse provenant de la collecte toujours plus massive des données. Dans cet optique, les prochaines versions viseront à continuer à étendre les types et structures de données traitables directement par l'outil et à paralléliser les algorithmes utilisés pour exploiter les capacités de traitement multi-coeurs et multi-machines actuellement disponibles.

Références

Asuncion, A. et D. Newman (2007). UCI machine learning repository.

Blockeel, H., L. De Raedt, et J. Ramon (1998). Top-Down Induction of Clustering Trees. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 55–63. Morgan Kaufmann.

Boullé, M. (2005). A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research 6*, 1431–1452.

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

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

Boullé, M. (2009). A parameter-free classification method for large scale learning. *Journal of Machine Learning Research 10*, 1367–1385.

Boullé, M. (2014). Towards automatic feature construction for supervised classification. In *ECML/PKDD 2014*, pp. 181–196. Springer-Verlag.

Džeroski, S. et N. Lavrač (2001). *Relational Data Mining*. Springer-Verlag New York, Inc.

Guyon, I., S. Gunn, M. Nikravesh, et L. Zadeh (2006). *Feature Extraction : Foundations And Applications*. Springer.

Knobbe, A. J., H. Blockeel, A. Siebes, et D. Van Der Wallen (1999). Multi-Relational Data Mining. In *Proceedings of Benelearn '99*.

Kramer, S., P. A. Flach, et N. Lavrač (2001). Propositionalization approaches to relational data mining. In S. Džeroski et N. Lavrač (Eds.), *Relational data mining*, Chapter 11, pp. 262–286. Springer-Verlag.

Krogel, M.-A. et S. Wrobel (2001). Transformation-based learning using multirelational aggregation. In *ILP*, pp. 142–155. Springer.

Liu, H. et H. Motoda (1998). *Feature Extraction, Construction and Selection : A Data Mining Perspective*. Kluwer Academic Publishers.

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

Summary

Khiops is an automatic supervised classification tool for mining large multi-tables databases. The predictive importance of input variables is evaluated by the mean of discretization models in the numerical case and of value grouping models in the categorical case. In the case of a multi-tables database, for exemple customers with their purchases, an analysis data table instances $\times$ variables is produced using automatic feature construction. The supervised classification model is a naive Bayes classifier, with variable selection and model averaging. The tool is designed for the analysis of large databases, with millions of instances, tens of thousands of variables and hundreds of millions of records in the secondary tables.