Sélection des variables informatives pour l'apprentissage supervisé multi-tables

Dhafer Lahbib*,** Marc Boullé*, Dominique Laurent**

*France Télécom R&D - 2, avenue Pierre Marzin, 23300 Lannion
dhafer.lahbib@orange-ftgroup.com
marc.boulle@orange-ftgroup.com
**ETIS-CNRS-Universite de Cergy Pontoise-ENSEA, 95000 Cergy Pontoise
dominique.laurent@u-cergy.fr

Résumé. Dans la fouille de données multi-tables, les données sont représentées sous un format relationnel dans lequel les individus de la table cible sont potentiellement associés à plusieurs enregistrements dans des tables secondaires en relation un-à-plusieurs. La plupart des approches existantes opèrent en transformant la représentation multi-tables, notamment par mise à plat. Par conséquent, on perd la représentation initiale naturellement compacte mais également on risque d'introduire des biais statistiques. Notre approche a pour objectif d'évaluer l'informativité des variables explicatives originelles par rapport à la variable cible dans le contexte des relations un-à-plusieurs. Elle consiste à résumer l'information contenue dans chaque variable par un tuple d'attributs représentant les effectifs des modalités de celle-ci. Des modèles en grilles multivariées sont alors employés pour qualifier l'information apportée conjointement par les nouveaux attributs, ce qui revient à une estimation de densité conditionnelle de la variable cible connaissant la variable explicative en relation un-à-plusieurs. Les premières expérimentations sur des bases de données artificielles et réelles montrent qu'on arrive à identifier les variables explicatives potentiellement pertinentes sur tout le domaine relationnel.

1 Introduction

Tandis que dans les méthodes de fouille de données classiques, les données sont stockées dans une seule table, la *Fouille de données mutli-tables* (en anglais, Multi-Relational Data Mining, MRDM) s'intéresse à l'extraction de connaissances à partir de bases de données relationnelles multi-tables (Knobbe et al., 1999). Typiquement, en MRDM les individus sont contenus dans une table *cible* en relation un-à-plusieurs avec des *tables secondaires*. En apprentissage supervisé, un *attribut cible* devrait être défini au sein de la table cible ce qui correspond à la *variable à expliquer* par analogie au cas mono-table.

Pour prendre en compte les relations un-à-plusieurs, la plupart des méthodes MRDM opèrent en transformant la représentation relationnelle : dans le paradigme de la Programmation Logique Inductive ILP (Džeroski, 1996), les données sont recodées sous la forme de

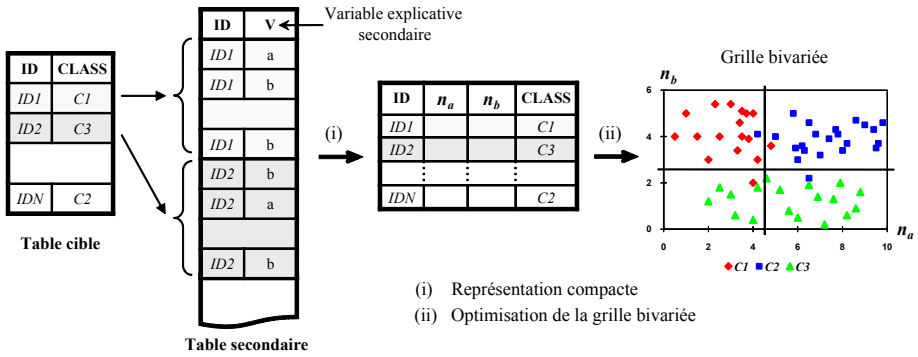


FIG. 1: Variable secondaire binaire

prédicats logiques, ce qui pose des problèmes de passage à l'échelle surtout sur de gros volumes de données (Blockeel et Sebag, 2003). D'autres méthodes dites par propositionalisation opèrent par mise-à-plat (Kramer et al., 2001). Elles cherchent à agréger l'information contenue dans les différentes tables, les transformant ainsi sous un format tabulaire classique par création de nouvelles variables. Par conséquent, non seulement on perd la représentation initiale naturellement compacte mais encore on risque d'introduire des biais statistiques, notamment à cause des dépendances qui peuvent exister entre les variables nouvellement créées.

Notre objectif dans cet article est d'étudier l'informativité d'une variable explicative située dans une table secondaire en relation *un-à-plusieurs* [1] avec la table cible. Elle peut être évaluée en mesurant la probabilité conditionnelle $P(Y \mid A)$, où Y est la variable cible et A est une variable explicative secondaire. Ceci peut être très utile pour une étape de sélection de variable de type filtre (Guyon et Elisseeff, 2003) ou encore pour proposer un algorithme de classification dans le contexte multi-tables nécessitant un prétraitement univarié telle que les méthodes de type Bayésien Naïf ou les arbres de décision.

Après avoir présenté dans la section 2 les principes de notre approche, nous décrivons dans la section 3 les résultat d'évaluation obtenus sur des données artificielles. Enfin, la section 4 conclut cet article.

2 Approche

Pour simplifier le problème, nous nous limitons dans cet article au cas le plus simple, celui d'une variable binaire qui ne peut prendre que deux valeurs a et b. Dans le cas mono-table, à chaque individu correspond une seule valeur pour la variable considérée. Dans le cas d'une relation un-à-plusieurs, un individu est décrit par une liste (éventuellement vide) de valeurs parmi a et b dont la taille varie d'un individu à l'autre. Évaluer la pertinence de la variable avec cette représentation est difficile. Une solution possible consiste à résumer la variable à l'aide d'une représentation équivalente plus compacte en introduisant deux nouvelles variables n_a et n_b qui

1. Dans le cas des relations un-à-un, on se ramène au cas mono-table. Précisons que pour des raisons de simplification, on se limite ici au premier niveau de relation : tables en relation directe avec la table cible.

représentent respectivement le nombre d'occurrences de a et de b[2]. Considérer ces deux nouvelles variables indépendamment l'une de l'autre ne permet pas de saisir toute l'information sur la variable initiale : l'information sur les ratios de chacune des valeurs ($\frac{n_a}{n_a+n_b}$ et $\frac{n_b}{n_a+n_b}$) ainsi que l'effectif total ($n_a + n_b$) seront perdus. La solution que nous proposons consiste à considérer les deux variables n_a et n_b conjointement. Ainsi toute l'information de la variable initiale est préservée. La probabilité $P(Y \mid A)$ sera alors équivalente à $P(Y \mid n_a, n_b)$. Pour qualifier l'information prédictive contenue dans la paire de variables numériques n_a et n_b, nous utilisons des modèles en grille de données selon l'approche MODL appliquée au cas bivarié (Boullé, 2007). La méthode consiste à discrétiser chacune des deux variables numériques en intervalles. Les individus sont alors partitionnés en une grille, dont les cellules sont définies par des paires d'intervalles. La distribution de la variable à expliquer dans chaque cellule se déduit à partir du produit cartésien des deux partitions univariées qui répartit les individus sur la grille (cf. figure 1). C'est une représentation interprétable puisqu'elle permet de voir la distribution des individus sur la grille en faisant varier conjointement les deux variables.

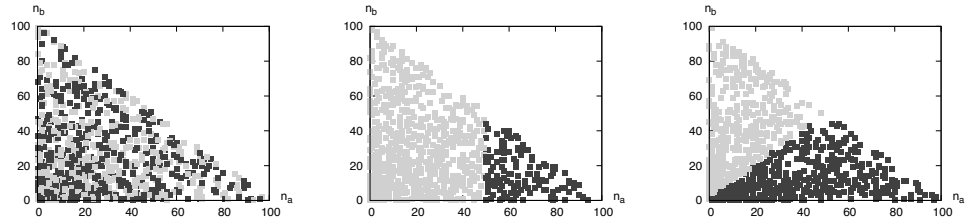
Dans l'approche MODL, le problème de partitionnement de deux variables numériques est transposé en un problème de sélection de modèles. La meilleure grille de données est choisie selon une approche MAP (maximum a posteriori), qui consiste à maximiser la probabilité d'un modèle en grille connaissant les données $p(\text{Modèle}|\text{Données})$. En appliquant la formule de Bayes, le problème se ramène à maximiser $p(\text{Modèle})p(\text{Données}|\text{Modèle})$. Notons par N le nombre d'individus de l'échantillon (nombre d'enregistrements de la table cible), J le nombre de valeurs de la variable à expliquer et $N_{i_a i_b .}$ le nombre d'individus de la cellule (i_a, i_b). Les paramètres d'un partitionnement particulier sont le nombre d'intervalles I_a et I_b, les bornes des intervalles $\{N_{i_a ..}\}_{1 \leq i_a \leq I_a}$ et $\{N_{.i_b .}\}_{1 \leq i_b \leq I_b}$ et les effectifs de la variable à expliquer $\{N_{i_a i_b j}\}_{1 \leq i_a \leq I_a, 1 \leq i_b \leq I_b}$ par cellule (i_a, i_b). Un modèle de partitionnement bivarié est donc entièrement défini par les paramètres $\{I_a, I_b, \{N_{i_a ..}\}, \{N_{.i_b .}\}, \{N_{i_a i_b j}\}\}$. La distribution a priori des modèles $p(\text{Modèle})$ et la vraisemblance des données $p(\text{Données}|\text{Modèle})$ sont calculées analytiquement en exploitant le caractère discret de la famille de modèles, et en adoptant des hypothèses faiblement informatives sur les données. La meilleure grille est celle qui minimise le critère de l'equation 1 (Boullé, 2007).

$$\begin{aligned} &\log N + \log N + \log\left(C_{I_a-1}^{N+I_a-1}\right) + \log\left(C_{I_b-1}^{N+I_b-1}\right) \\ &+ \sum_{i_a=1}^{I_a} \sum_{i_b=1}^{I_b} \log\left(C_{J-1}^{N_{i_a i_b .}+J-1}\right) + \sum_{i_a=1}^{I_a} \sum_{i_b=1}^{I_b} \log \frac{N_{i_a i_b .}!}{N_{i_a i_b 1}! N_{i_a i_b 2}! \ldots N_{i_a i_b J}!} \end{aligned} \quad (1)$$

Les cinq premiers termes représentent la probabilité *a priori* du modèle : le choix des nombres d'intervalles, des bornes des intervalles, et de la distribution de la variable à expliquer dans chaque cellule de la grille. Le dernier terme correspond à la vraisemblance.

Le critère d'évaluation du partitionnement bivarié est optimisé en partant d'une solution initiale aléatoire et en alternant les optimisations partielles par variable. Les optimisations partielles sont effectuées grâce à une heuristique gloutonne ascendante décrite dans (Boullé, 2006). Ce critère permet d'effectuer une sélection de variables de type filtre en classant celles-ci par valeur de critère décroissante.

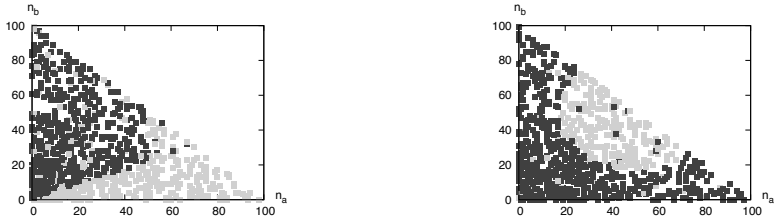
2. Cette représentation est bien équivalente puisqu'on peut retrouver toute l'information initiale (nous supposons que l'ordre des éléments dans la table secondaire n'a pas d'importance et nous ne considérons qu'une seule variable à la fois).



(a) Motif "Aléatoire" :
$P(c=0) = P(c=1) = 0.5$

(b) Motif basé sur les "Effectifs" :
si $n_a \geq$ seuil alors $c = 1$
si $n_a <$ seuil alors $c = 0$

(c) Motif basé sur les "Proportions" :
si $\frac{n_a}{n_a+n_b} \geq R$ alors $c = 1$
si $\frac{n_a}{n_a+n_b} < R$ alors $c = 0$



(d) Motif combinant "Effectifs et Proportions" :
si $n_a \geq N$ et $\frac{n_a}{n_a+n_b} \leq R$ alors $c = 0$ avec 5% de bruit
sinon $c = 1$ avec 5% de bruit

(e) "Motif complexe" :
si $(n_a - M_1)^2 + (n_b - M_2)^2 \geq R$ alors $c = 1$ avec 5% de bruit
si $(n_a - M_1)^2 + (n_b - M_2)^2 < R$ alors $c = 0$ avec 5% de bruit

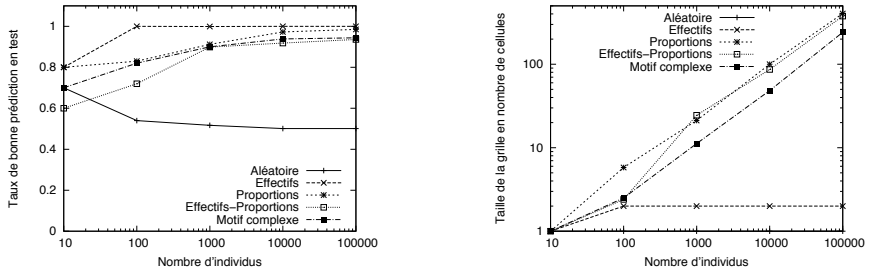
FIG. 2: Motifs et diagrammes de dispersion des bases de données artificielles générées

3 Expérimentations

Pour évaluer l'intérêt de notre approche, des tests ont été effectués sur des données artificielles [3]. Les bases de données présentent une structure relationnelle composée de deux tables : une table cible en relation un-à-plusieurs avec une table secondaire contenant une seule variable explicative binaire. Les paramètres du générateur de données sont le nombre d'individus (enregistrements de la table cible) et le nombre maximal d'enregistrements reliés à chaque individu dans la table secondaire. La valeur de la variable à expliquer (booléenne) est générée selon un motif dans la variable secondaire. La figure 2 illustre les diagrammes de dispersion des bases de données artificielles générées ainsi que les motifs utilisés (ici le nombre d'enregistrements secondaires par individu est aléatoire entre 0 et 100).

Pour chaque base de données on applique l'approche décrite dans la section 2 afin d'obtenir la grille bivariée optimale correspondante. Cette grille est utilisée comme une table de décision. Pour classifier un individu en test. On recherche la cellule de la grille associée aux valeurs de l'individu pour la paire de variables, et on prédit la valeur de la classe majoritaire de la cellule (d'après les effectifs collectés en apprentissage). La pertinence d'une variable explicative secondaire est évaluée en se basant sur le taux de bonne prédiction et l'aire sous la courbe ROC (AUC) en test du classifieur en grille correspondant. Pour cela, nous appliquons une validation

3. D'autres tests sur la base réelle STULONG ont été également faits, mais non commentés ici faute de place.



(a) Test Accuracy en fonction du nombre d'individus (b) Taille de la grille en fonction du nombre d'individus

FIG. 3: Résultats obtenus sur les données artificielles de la figure 2

croisée stratifiée d'ordre 10. La taille de la grille optimale informe sur la complexité du motif éventuel dans la variable explicative.

La figure 3a illustre les résultats de classification pour les 5 bases de données artificielles et ceci pour un nombre d'individus variable. Comme première constatation, les résultats montrent que la méthode permet de détecter facilement un motif totalement aléatoire. L'absence de l'information prédictive dans la grille bivariée se matérialise par une discrétisation en un seul intervalle (figure 3b) et un taux de bonne prédiction de l'ordre de 50%. Le modèle nul est d'autant plus confirmé qu'il y a d'individus. La méthode permet également de détecter des motifs plus ou moins complexes. La figure 3a montre que les performances en classification s'améliorent avec le nombre d'individus dans la base et qu'avec suffisamment d'individus, les taux de bonne prédiction atteignent approximativement les performances théoriques (vu l'existence de bruit). La taille de la grille bivariée varie selon la complexité du motif : pour un motif assez simple basé sur les effectifs, la grille est toujours composée de deux cellules ; pour des motifs plus complexes la taille de la grille augmente avec le nombre d'individus, permettant ainsi d'approximer finement le motif.

4 Conclusion

Dans cet article, nous avons proposé une approche pour évaluer la pertinence d'une variable explicative dans le contexte de l'apprentissage supervisé à partir des données relationnelles. La méthode consiste à changer la représentation relationnelle initiale en une représentation tabulaire équivalente. Les attributs générés représentent les effectifs des valeurs de la variable initiale. Un modèle de discrétisation en intervalles de chacun de ces attributs est alors généré, ce qui induit une partition multivariée. Cette partition permet de qualifier l'information apportée conjointement par tous les nouveaux attributs sur la variable cible. L'information sur la variable explicative initiale n'est pas perdue ce qui est alors équivalent à évaluer la pertinence de celle-ci. Un critère d'évaluation est proposé dans le cas d'une variable secondaire binaire pour évaluer chaque partition bivariée au moyen d'une approche Bayesienne. Nous avons évalué l'approche sur des bases de données artificielles et réelles. Les premiers résultats sur des variables explicatives binaires montrent que le critère d'évaluation permet de sélectionner des variables fortement informatives. Des travaux futurs sont envisagés pour proposer les procédures d'optimisation efficaces de ce critère dans le cas de variables catégorielles (éven-

tuellement avec un nombre important de valeurs) et continues, et d'intégrer les prétraitements des variables un-à-plusieurs dans des classifieurs de type Bayésien Naïf ou arbre de décision.

Références

Blockeel, H. et M. Sebag (2003). Scalability and efficiency in multi-relational data mining. *ACM SIGKDD Explorations Newsletter 5*(1), 17.

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

Boullé, M. (2007). Une méthode optimale d'évaluation bivariée pour la classification supervisée. *Extraction et gestion des connaissances (EGC'2007)*, 461–472.

Džeroski, S. (1996). *Inductive logic programming and knowledge discovery in databases*, pp. 117–152. Menlo Park, CA, USA : American Association for Artificial Intelligence.

Guyon, I. et A. Elisseeff (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research 3*, 1157–1182.

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*, Chapter 11, pp. 262–286. New York, NY, USA : Springer-Verlag.

Summary

In multi-relational data mining, data is represented in a relational form where the individuals of the target table are potentially related to several records in secondary tables in one-to-many relationship. To cope with this one-to-many setting, most of the existing approaches try to transform the multi-tables representation, namely by propositionalisation, thereby losing the naturally compact initial representation and eventually introducing statistical bias. Our approach aims to directly evaluate the informativness of the original input variables over the relational domain w.r.t. the target variable. The idea is to summarize for each individual the information contained in the non target table variable by a features tuple representing the cardinalities of the initial modalities. Multivariate grid models have been used to qualify the joint information brought by the new features, which is equivalent to estimating the conditional density of the target variable given the input variable in non target table. Preliminary experiments on artificial and real data sets show that potentially relevant one-to-many variables could be found.