Un Critère d'évaluation pour la construction de variables à base d'itemsets 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 le contexte de la fouille de données multi-tables, les données sont représentées sous un format relationnel dans lequel les individus de la table cible sont potentiellement liés à plusieurs enregistrements dans des tables secondaires en relation un-à-plusieurs. Dans cet article, nous proposons un Framework basé sur des itemsets pour la construction de variables à partir des tables secondaires. L'informativité de ces nouvelles variables est évaluée dans le cadre de la classification supervisée au moyen d'un critère régularisé qui vise à éviter le sur-apprentissage. Pour ce faire, nous introduisons un espace de modèles basés sur des itemsets dans la table secondaire ainsi qu'une estimation de la densité conditionnelle des variables construites correspondantes. Une distribution a priori est définie sur cet espace de modèles, pour obtenir ainsi un critère sans paramètres permettant d'évaluer la pertinence des variables construites. Des expérimentations préliminaires montrent la pertinence de l'approche.

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, une *variable cible* devrait être définie au sein de la table cible. La nouveauté en MRDM est de considérer les variables se trouvant dans les tables secondaires (*variables secondaires*) pour prédire la classe. Plusieurs solutions ont été proposée dans la littérature, notamment la Programmation Logique Inductive PLI (Džeroski, 1996) qui utilise le formalisme logique ou encore la propositionalisation qui opèrent par mise à plat afin de pouvoir utiliser un classifieur monotable classique (Kramer et al., 2001).

Dans cet article, nous introduisons un espace de modèles basé sur des itemsets de variables secondaires. Ces itemsets permettent de construire de nouvelles variables binaires dans les tables secondaires. Ensuite nous évaluons la pertinence de ces variables pour la tâche de classification supervisée. Afin de prendre en compte le risque de sur-apprentissage, qui augmente

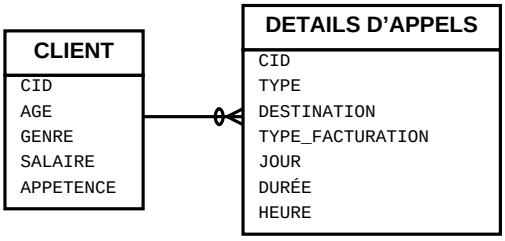

FIG. 1 – *Schéma relationnel d'une base de données CRM. La table cible CLIENT est en relation un-à-plusieurs avec la table secondaire DÉTAILS D'APPELS.*

considérablement avec le nombre potentiel des variables construites, nous définissons une distribution *a priori* sur cet espace de modèles ainsi qu'une estimation de la densité conditionnelle de ces variables. Nous obtenons ainsi un critère sans paramètres pour évaluer la pertinence des variables ainsi construites. Nous illustrons notre approche à travers l'exemple ci-dessous :

Exemple 1. La figure 1 illustre un extrait du schéma relationnel d'une base de données CRM. Le problème est, par exemple, de prédire l'appétence des clients à un certain produit. La variable cible est la variable booléenne APPÉTENCE qui indique si le client est susceptible de commander ce produit. Pour ce faire, nous considérons des itemsets π de variables secondaires constitués d'une conjonction d'expressions de la forme $(x \in S_x)$, où x est une variable secondaire de la table DÉTAILS D'APPEL et S_x est un groupe de valeurs si x est catégorielle, ou un intervalle si x est numérique. Si on suppose que les variables secondaires JOUR et DESTINATION sont catégorielles et que la variable DURÉE est numérique, alors $\pi : (JOUR \in \{Samedi,\ Dimanche\}) \wedge (DESTINATION \in \{International\}) \wedge (DUR\acute{E}E \in [10; 30])$ est un itemset.

Un itemset permet de construire une nouvelle variable binaire A_π dans la table secondaire selon que les enregistrements de cette table sont couverts ou non par l'itemset. Le critère d'évaluation proposé se décompose en la somme de deux termes : (i) un coût de codage évaluant la construction de l'itemset π, et (ii) un critère qui estime la pertinence de la variable A_π par rapport à la variable cible, qui exploite l'approche d'évaluation des variables secondaires binaires introduite dans (Lahbib et al., 2011).

Le reste de cet article est organisé comme suit. La partie 2 introduit l'espace des variables construites à base d'itemsets de variables secondaires et présente son critère d'évaluation. Dans la partie 3 nous évaluons la méthode sur un jeu de données réelles. Enfin, la partie 4 conclut cet article et discute sur des travaux futurs.

2 Construction de variables à base d'itemsets

En se basant sur le modèle de classification de (Gay et Boullé, 2012), un itemset π est une conjonction d'expressions de la forme $(x \in S_x)$, où x est une variable secondaire et S_x est soit un groupe de valeurs si x est catégorielle, soit un intervalle si x est numérique. À chaque itemset π nous associons une variable secondaire binaire A_π. Celle-ci est évaluée à « vrai » pour les enregistrements secondaires couverts par π, et à « faux » sinon. Notre espace

de modèles est donc celui de tous les itemsets. Afin d'appliquer une démarche Bayésienne, nous définissons d'abord une distribution *a priori* sur cet espace. Pour cela, nous introduisons les notations suivantes :

- N : nombre d'individus de l'échantillon (nombre d'enregistrements de la table cible)
- J : nombre de valeurs de la variable à expliquer (connu)
- N_s : nombre de lignes de la table secondaire
- m : nombre de variables (numériques et catégorielles) dans la table secondaire
- $X = \{x_1, \ldots, x_k\}$: l'ensemble des k variables secondaires qui constituent l'itemset
- X_{cat} (resp. X_{num}) : ensemble des variables catégorielles (resp. numériques) qui apparaissent dans l'itemset ($X = X_{cat} \cup X_{num}$)
- V_x : nombre de valeurs de la variable secondaire catégorielle $x \in X$ ($V_x = |\text{dom}(x)|$)
- I_x : nombre d'intervalles (resp. groupes de valeurs) de la variable secondaire $x \in X$ numérique (resp. catégorielle)

A priori hiérarchique d'un itemset de variables secondaires. Nous utilisons l'*a priori* hiérarchique défini ci-dessous. Soulignons qu'une distribution uniforme est utilisée à chaque étage [1] de la hiérarchie des paramètres des modèles.

1. le nombre k de variables secondaires qui constituent l'itemset est uniformément distribué entre 0 et m.
2. pour un nombre de variables k, chaque sous-ensemble de k variables qui constituent l'itemset est équiprobable dans un tirage avec remise.
3. pour une variable secondaire catégorielle qui figure dans l'itemset, le nombre de groupes est nécessairement 2 ($I_x = 2$).
4. pour une variable secondaire numérique qui figure dans l'itemset, le nombre d'intervalles est soit 2, soit 3 de façon équiprobable.
5. pour une variable secondaire numérique (respectivement, catégorielle), et pour un nombre d'intervalles (respectivement, nombre de groupes) donné, toutes les partitions en I_x intervalles (respectivement, en I_x groupes de valeurs) sont équiprobables.
6. pour une variable secondaire catégorielle x appartenant à l'itemset, le choix du groupe de valeurs i_x sur lequel porte la condition est équiprobable.
7. pour une variable secondaire numérique x appartenant à l'itemset, si la variable est discrétisée en deux intervalles, le choix de celui sur lequel porte la condition est équiprobable. Lorsqu'il y a 3 intervalles, celui qui figure dans l'itemset est nécessairement l'intervalle du milieu.

En utilisant la définition de l'espace de modèles ainsi que sa distribution a priori, le coût de construction $\mathcal{C}_c(A_\pi)$ d'un itemset π est donné dans l'équation 1.

$$\begin{aligned}\mathcal{C}_c(A_\pi) = \quad & \log(m+1) + \log\left(\begin{array}{c} m+k-1 \\ k \end{array}\right) + \sum_{x \in X_{cat}} \left(\log\left(\mathcal{S}(V_x, 2)\right) + \log 2\right) \\ & + \sum_{x \in X_{num}} \left(\log 2 + \log\left(\begin{array}{c} N_s + I_x - 1 \\ I_x - 1 \end{array}\right) + \log\left(1 + \mathbb{1}_{\{2\}}(I_x)\right)\right)\end{aligned} \tag{1}$$

1. Cela ne signifie pas que l'*a priori* hiérarchique est un *a priori* uniforme sur l'espace des itemsets, ce qui serait équivalent à une approche par maximum de vraisemblance.

Les deux premiers termes de l'équation 1 correspondent au choix du nombre de variables secondaires qui apparaissent dans l'itemset ainsi que le choix de ces variables parmi toutes les variables de la table secondaire. Le troisième termes représente le choix des partitions des valeurs des variables secondaires catégorielles ainsi que le choix des groupes impliqués dans l'itemset où $\mathcal{S}$ dénote le nombre de Stirling de deuxième espèce. La troisième ligne correspond au choix de la discrétisation des variables secondaires numériques ainsi que les intervalles sur lesquels portent les conditions de l'itemset. Le critère de l'équation 1 est un log négatif de probabilités, ce qui exprime une longueur de codage Shannon (1948). $\mathcal{C}_c\left(A_\pi\right)$ peut être donc interprété comme un coût de codage de l'itemset π. Par ailleurs, il peut être vu comme un coût de construction de la variable A_π associée à π.

Évaluation d'une variable secondaire binaire. La variable A_π associée à π est une variable binaire construite dans la table secondaire. Lahbib et al. (2011) fournissent une approche d'estimation de densité de probabilité conditionnelle d'une telle variable vis-à-vis de la variable cible, ainsi qu'un critère permettant d'évaluer sa pertinence $\mathcal{C}_e\left(A\right)$. Ce critère est rappelé dans l'équation 2.

$$\begin{aligned}\mathcal{C}_e\left(A\right) = \quad & \log N + \log N + \log \left(\begin{array}{c} N + I_a - 1 \\ I_a - 1 \end{array} \right) + \log \left(\begin{array}{c} N + I_b - 1 \\ I_b - 1 \end{array} \right) \\ & + \sum_{i_a=1}^{I_a} \sum_{i_b=1}^{I_b} \log \left(\begin{array}{c} N_{i_a i_b \cdot} + J - 1 \\ J - 1 \end{array} \right) + \sum_{i_a=1}^{I_a} \sum_{i_b=1}^{I_b} \log \frac{N_{i_a i_b \cdot}!}{N_{i_a i_b 1}! N_{i_a i_b 2}! \ldots N_{i_a i_b J}!}\end{aligned} \tag{2}$$

Les détails sur ce critère ainsi que l'algorithme d'optimisation peuvent être trouvés dans (Lahbib et al., 2011).

Critère d'évaluation global. Le critère d'évaluation global d'un itemset π s'obtient en remplaçant chaque terme par son expression dans l'équation 3.

$$\mathcal{C}_r\left(A_\pi\right) = \mathcal{C}_e\left(A_\pi\right) + \mathcal{C}_c\left(A_\pi\right) \tag{3}$$

Le coût de construction agit comme un terme de régularisation afin de prévenir le risque de sur-apprentissage lié au grand nombre d'itemsets potentiellement considérés. Les variables secondaires A_π construites sur la base d'itemsets complexes, avec un nombre important de variables dans l'itemset, sont pénalisées par rapport à des variables construites plus simples. Soit $\pi_\emptyset$ un itemset vide, ne contenant aucune variable secondaire, où aucun enregistrement secondaire n'est couvert par l'itemset. Le coût d'évaluation global $\mathcal{C}_r\left(A_{\pi_\emptyset}\right)$ de l'itemset vide est :

$$\begin{aligned}\mathcal{C}_r\left(A_{\pi_\emptyset}\right) & = \log(m+1) + 2\log N + \log \tfrac{N!}{N_1! N_2! \ldots N_J!} \\ & = N.Ent(Y) + O(\log N)\end{aligned} \tag{4}$$

où $Ent(Y)$ est l'entropie de la variable cible Y, et N_j $(1 \leq j \leq J)$ est le nombre d'individus ayant pour classe la valeur j. Par conséquent, tout itemset π ayant un coût global supérieur à celui de l'itemset vide peut être ignoré, puisqu'il apporte moins d'information que la variable cible seule.

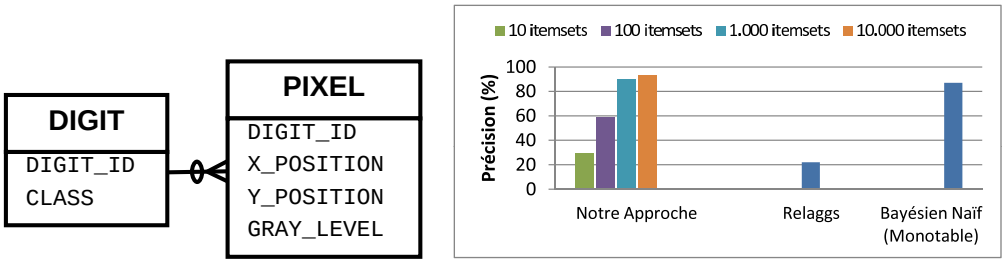

FIG. 2 – *Schéma relationnel et résultats de la base de données Digits.*

3 Expérimentations

Nous avons évalué notre approche en utilisant la bases de données Digits (Lecun et al., 1998). Il s'agit de classer des images représentant des chiffres manuscrits de 0 à 9. Ces données initialement à plat [2] ont été reformatées afin d'obtenir un schéma relationnel (figure 2) constitué de deux tables : la table cible DIGIT et la table secondaire PIXEL décrivant les pixels qui composent chaque image. Cette dernière est décrite par trois variables secondaires : X_POSITION, Y_POSITION et GRAY_LEVEL qui représentent la position du pixel en abscisse et en ordonnée dans l'image originale ainsi que son niveau de gris.

Nous générons aléatoirement des itemsets basés sur des partitionnements (discrétisation dans le cas numérique et groupement de valeurs dans le cas catégoriel) des variables secondaires qui le constituent en 2, 4 et 8 partiles. Un classifieur Bayésien Naïf est employé en exploitant l'estimation de la densité de probabilité conditionnelle de la variable construite comme décrit dans (Lahbib et al., 2011).

La figure 2 illustre les performances de classification (Précision) du Bayésien Naïf utilisant les variables générées et ceci pour différents nombres d'itemsets (10, 1000, 1000 et 10 000). Ces résultats sont comparés à ceux obtenus avec le système Relaggs (Krogel et Wrobel, 2001). Relaggs est une méthode de propostionalisation qui consiste à générer pour chaque variable secondaire plusieurs agrégats (les effectifs, la somme, la moyenne, le min, le max, l'écart type,...). Ces agrégats sont ensuite ajoutés à la table cible et un Bayésien Naïf classique est employé. Nous reportons également les performances obtenues avec un Bayésien Naïf utilisant la représentation monotable initiale. On peut constater que notre approche dépasse largement Relaggs et ceci pour tous les nombres d'itemsets générés. Par ailleurs, avec suffisamment d'itemsets, notre approche atteint des performances comparables à celles d'un Bayésien Naïf qui utilise la représentation à plat.

4 Conclusion

Dans cet article, nous avons proposé une approche de prétraitement multivarié des variables secondaires dans le contexte de la classification de données multi-tables. La méthode consiste à construire de nouvelles variables à partir d'itemsets de variables secondaires. La pertinence

2. http://yann.lecun.com/exdb/mnist/

de cette nouvelle variable est évaluée en utilisant un modèle en grille de données bivariée, qui fournit un estimateur régularisé de la probabilité conditionnelle de la variable cible.

Afin d'éviter le sur-apprentissage, nous avons appliqué une approche Bayésienne de sélection de modèles pour la construction des itemsets ainsi que pour le modèle d'évaluation de la densité conditionnelle. Nous obtenons ainsi un critère analytique exact pour l'estimation de la probabilité *a posteriori* de la variable construite. Nous envisageons dans des travaux futurs de fournir des heuristiques performantes de recherche pour explorer l'espace des variables construites et de garder les plus pertinentes avec l'estimation de leur probabilité conditionnelle.

Références

Džeroski, S. (1996). Inductive logic programming and knowledge discovery in databases. In *Advances in knowledge discovery and data mining*, pp. 117–152. AAAI.

Gay, D. et M. Boullé (2012). A Bayesian approach for classification rule mining in quantitative databases. In *ECML PKDD '2012*.

Knobbe, A. J., H. Blockeel, A. Siebes, et D. Van Der Wallen (1999). Multi-Relational Data Mining. In *Proceedings of 9th annual ML conference of Belgium and The Netherlands*.

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

Krogel, M.-A. et S. Wrobel (2001). Transformation-based learning using multirelational aggregation. *Proceedings of the 11th International Conference on ILP*, 142–155.

Lahbib, D., M. Boullé, et D. Laurent (2011). Sélection des variables informatives pour l'apprentissage supervisé multi-tables. In *EGC' 2011*.

Lecun, Y., L. Bottou, Y. Bengio, et P. Haffner (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*(11), 2278–2324.

Shannon, C. (1948). A mathematical theory of communication. Technical report. *Bell systems technical journal*.

Summary

In multi-relational data mining, data are represented in a relational form where the individuals of the target table are potentially related to several records in secondary tables through one-to-many relationship. In this paper, we introduce an itemset based framework for constructing variables from secondary tables. The relevance of these new variables is evaluated in the context of supervised classification using a regularized criterion in order to avoid overfitting. To this end, we introduce a space of itemset based models in the secondary table and a conditional density estimation of the related constructed variables. A prior distribution is defined on this model space, thereby obtaining a parameter-free criterion to assess the relevance of the constructed variables. Preliminary experiments show the relevance of the approach.