Prétraitement Supervisé des Variables Numériques pour la Fouille de Données 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. Le prétraitement des variables numériques dans le contexte de la fouille de données multi-tables diffère de celui des données classiques individu-variable. La difficulté vient principalement des relations un-à-plusieurs où les individus de la table cible sont potentiellement associés à plusieurs enregistrements dans des tables secondaires. Dans cet article, nous décrivons une méthode de discrétisation des variables numériques situées dans des tables secondaires. Nous proposons un critère qui évalue les discrétisations candidates pour ce type de variables. Nous décrivons un algorithme d'optimisation simple qui permet d'obtenir la meilleure discrétisation en intervalles de fréquence égale pour le critère proposé. L'idée est de projeter dans la table cible l'information contenue dans chaque variable secondaire à l'aide d'un vecteur d'attributs (un attribut par intervalle de discrétisation). Chaque attribut représente le nombre de valeurs de la variable secondaire appartenant à l'intervalle correspondant. Ces attributs d'effectifs sont conjointement partitionnés à l'aide de modèles en grille de données afin d'obtenir une meilleure séparation des valeurs de la classe. Des expérimentations sur des jeux de données réelles et artificielles révèlent que l'approche de discrétisation permet de découvrir des variables secondaires pertinentes.

1 Introduction

La plupart des algorithmes de fouille de données existants se basent sur une représentation des données de la forme attribut-valeur. Dans ce format dit à plat, chaque enregistrement représente un individu et les colonnes représentent les variables décrivant ces individus. Dans les applications réelles, les données présentent souvent une structure interne difficile à exprimer sous un format tabulaire. Cette structure peut être naturellement décrite à l'aide du formalisme relationnel dans lequel chaque objet est relié à un ou plusieurs enregistrements dans d'autres tables (tables secondaires) à l'aide de relations de clés étrangères.

Exemple 1. *Prenons par exemple un problème CRM (Customer Relationship Management). La Figure 1 représente un extrait du schéma relationnel d'une base de données fictive d'une*

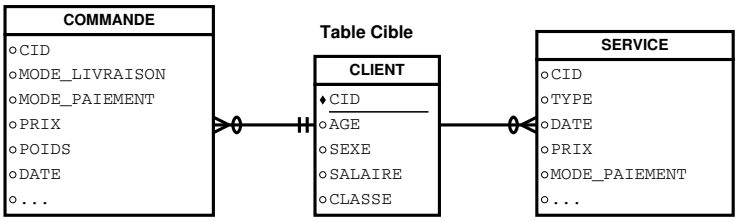


FIG. 1 – *Schema relationnel d'une base de données CRM fictive*

application CRM. Le problème est par exemple, d'identifier les clients susceptibles de s'intéresser à un certain produit ou service, ce qui revient à un problème de classification de clients. La variable cible correspond à l'attribut CLASSE qui dénote si le client a déjà commandé un produit particulier. La table CLIENT est en relation un-à-plusieurs *avec des* tables secondaires *(COMMANDE, SERVICE).*

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 (Knobbe et al., 1999). Les données multi-tables violent deux hypothèses majeures des algorithmes de classification classiques. D'une part, ces derniers supposent que les instances (objets à classer) possèdent une structure "homogène", notamment, une seule valeur par attribut et ceci pour un nombre fixe de variables. Cependant, dans le cas multi-tables, pour une variable située dans une table secondaire, un individu peut avoir une liste de valeurs (éventuellement vide) dont la taille est variable. D'autre part, les données relationnelles ne sont pas i.i.d (Indépendamment Identiquement Distribués). à cause des dépendances entre les tables.

Il s'agit d'une nouvelle famille de techniques d'extraction de connaissances à partir de bases de données relationnelles multi-tables. Le point commun entre la majorité de ces techniques est qu'elles ont besoin d'effectuer une transformation de la représentation relationnelle : dans le paradigme de la Programmation Logique Inductive ILP (Džeroski, 1996), les données sont recodées sous la forme de formules logiques, ce qui pose des problèmes de passage à l'échelle surtout sur de gros volumes de données (Blockeel et Sebag, 2003). D'autre méthodes 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 en créant 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.

Alors que le prétraitement de variables est au centre des systèmes de fouille de données usuels (mono-table), cette étape a reçu peu d'attention dans le cadre de la fouille de données multi-tables. Le prétraitement, y compris la sélection de variables et la discrétisation des variables numériques peuvent non seulement améliorer les performances de classification mais encore réduire l'espace de recherche en MRDM.

Exemple 2. *Dans le problème de la Figure 1, prédire si un client pourrait être intéressé par un produit ne dépend pas uniquement des informations sur ce client. Les données concernant les autres produits qu'il a commandés peuvent potentiellement être très informatives. Des attributs comme le type, le prix du produit peuvent présenter des corrélations avec la variable*

cible ce qui est très utile pour prédire la valeur de celle-ci. Le prétraitement, notamment la la discrétisation de ce genre de variable pose plusieurs problèmes, en particulier, à cause des relations un-à-plusieurs.

À notre connaissance, peu de travaux dans la littérature se sont confrontés au problème de discrétisation des variables dans le cadre de la fouille de données multi-tables. Dans ce contexte, les approches de discrétisation diffèrent selon 2 axes : (i) si elles font usage de l'étiquette de la classe et (ii) si elles considèrent les relations un-à-plusieurs lors du calcul des points de coupure. Les méthodes de discrétisation en intervalles de *largeur égale* et de *fréquence égale* sont non supervisées et calculent les frontières, indépendamment de la structure multi-tables. La première génère k intervalles de taille égale, la seconde discrétise la variable de façon à ce que chaque intervalle ait approximativement le même nombre de valeurs. Pour tenir compte des relations un-à-plusieurs, Knobbe (2004) part de l'idée que les valeurs numériques d'une variable secondaire ne sont pas toutes pertinentes si on veut les comparer à un certain seuil : il suffit de considérer les valeurs minimale et maximale liées à chaque individu. Ainsi, il propose simplement de calculer un histogramme de fréquence égale en considérant seulement le minimum et le maximum pour chaque individu (ce qui produit respectivement deux discrétisations). Knobbe et Ho (2005) ont proposé une méthode de discrétisation à *intervallles de poids égal*. Ils reprennent l'idée proposée par Van Laer et al. (1997) : les individus reliés à un nombre plus important d'individus dans la table secondaire ont une influence plus grande sur le choix des frontières de discrétisation puisqu'ils contribuent avec plus de valeurs numériques. Afin de compenser cet impact, les valeurs numériques sont pondérées par l'inverse de la taille des sacs d'enregistrements auxquels ils appartiennent. Au lieu de produire des intervalles de fréquence égale, les points de coupure sont calculés de telle sorte que des intervalles à poids égal peuvent être obtenus. Toutes les méthodes ci-dessus sont non supervisées car elles n'utilisent pas les étiquettes de classe. Afin de prendre en compte à la fois l'information de variable cible et l'association un-à-plusieurs, Alfred (2009) propose une modification de la discrétisation à base d'entropie introduite par Fayyad et Irani (1993). Les mesures d'entropie sont calculées à partir des tables secondaires en propageant la variable cible vers celles-ci. Dans ce cas, les données ne sont plus i.i.d. (Indépendamment Identiquement Distribuées) puisqu'un enregistrement ne représente plus une instance et un individu peut apparaitre plusieurs fois.

Dans cet article, nous nous intéressons au problème de prétraitement d'une variable numérique située dans une table secondaire en relation un-à-plusieurs avec la table cible [1]. Nous proposons de discrétiser les valeurs d'une variable secondaire A et d'utiliser un critère d'optimisation afin d'obtenir le meilleur partitionnement de façon à ce que les valeurs de la classe soient maximalement différenciées. L'idée est d'utiliser les modèles en grilles de données pour estimer la probabilité conditionnelle $P(Y \mid A)$, où Y est la variable cible. Ce prétraitement univarié étendu au contexte multi-tables peut être très utile pour une étape de sélection de variables (Guyon et Elisseeff, 2003) ou encore comme préparation pour des classifieurs comme le Bayésien Naïf ou les arbres de décision.

Le reste de ce manuscrit est organisé comme suit : la partie 2 décrit notre approche. Dans la partie 3 nous évaluons la méthode sur des jeux de données artificielles et réelles. Enfin, la partie 4 conclut cet article et discute sur des travaux futurs.

1. les associations un-à-un sont équivalentes au cas mono-table. Pour simplifier le problème, nous nous limitons ici au premier niveau de relation : tables en relation directe avec la table cible.

2 Prétraitement des Variables Secondaires

Dans cette partie, nous décrivons comment une variable numérique appartenant à une table secondaire peut être discrétisée de façon supervisée.

2.1 Illustration de l'Approche

Prenons le cas le plus simple, celui d'une variable binaire qui ne peut prendre que deux valeurs v_1 et v_2. Dans ce cas, chaque individu est décrit par un multi-ensemble de valeurs parmi v_1 et v_2 [2]. Étant donné un individu, tout ce qu'on a besoin de savoir sur la variable secondaire est le nombre de v_1 et de v_2 dans la liste d'enregistrements reliée à cet individu (nous les notons respectivement n_1 et n_2). Ainsi, toute l'information de la variable initiale est préservée en considérant conjointement le couple (n_1, n_2). En utilisant cette représentation, la probabilité conditionnelle $P(Y \mid A)$ est alors équivalente à $P(Y \mid n_1, n_2)$.

Cette approche peut être généralisée aux variables numériques secondaires. L'idée est de discrétiser la variable numérique en K intervalles et de créer ensuite dans la table cible K nouvelles variables n_k ($1 \leq k \leq K$). Pour chaque individu, n_k désigne le nombre d'enregistrements reliés dans la table secondaire ayant une valeur appartenant au $k^{\text{ème}}$ intervalle. Comme dans le cas bivarié, évaluer $P(Y \mid A)$ est équivalent à évaluer $P(Y \mid (n_1, n_2, \cdots, n_K))$.

Les modèles en grille de données multi-variées sont des estimateurs non paramétriques de la probabilité conditionnelle de la classe sachant un ensemble de variables explicatives (Boullé, 2011). L'idée est de discrétiser conjointement les variables numériques n_k en intervalles. Ce partitionnement multi-varié définit une distribution des instances dans une grille de données à K-dimensions dont les cellules sont définies par des n-uplets d'intervalles. Par conséquent, notre objectif est de trouver la discrétisation multi-variée optimale qui maximise la séparation des classes. En d'autres termes, nous cherchons la grille optimale avec des cellules homogènes selon les valeurs de la variable cible.

Exemple 3. *Dans le contexte de l'exemple 2, considérons la variable secondaire "PRIX" dans la base de données de la Figure 1. Supposons que l'on discrétise cette variable en deux intervalles :* $]inf; 20]$ *and* $]20; sup[$*. Alors le prix est équivalent à la paire de variables* $\left(n_{]inf;20]}, n_{]20;sup[}\right)$ *où* $n_{]inf;20]}$ *(respectively* $n_{]20;sup[}$*) représente le nombre de commandes dont le prix est inférieur à 20 (respectivement supérieur à 20). Si on suppose que le prix est en corrélation avec la variable cible et que la discrétisation en deux intervalles est pertinente, les valeurs cibles peuvent être séparées facilement, en utilisant une grille similaire à celle de la Figure 2.*

La corrélation entre les cellules de la grille de données et les valeurs cibles permet de quantifier l'information de classification. La probabilité conditionnelle $P(Y \mid A)$ est évaluée localement dans chaque cellule. Par conséquent, des classifieurs comme le Bayesien Naïf ou les arbres de décisions peuvent facilement être utilisés. Par ailleurs, il est important de souligner que la grille de données fournit une représentation interprétable, puisqu'elle montre la distribution des individus en faisant varier conjointement les variables n_k. Chaque cellule peut être interprétée comme une règle de classification dans le contexte multi-tables.

2. Ceci est différent du cas mono-table, où à chaque individu correspond une seule valeur pour la variable considérée.

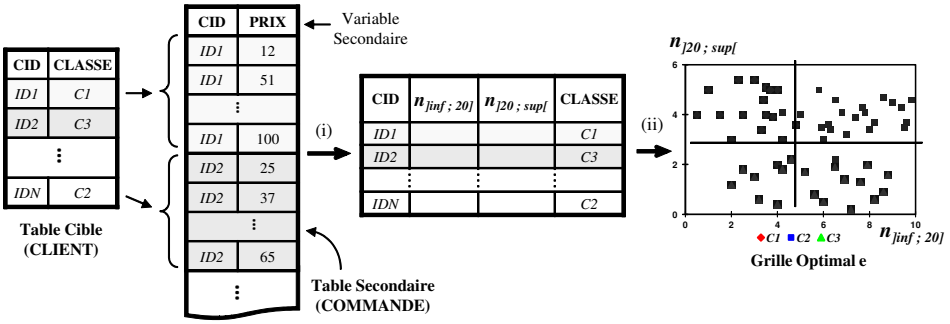


FIG. 2 – *Illustration de l'Approche*

Par exemple, la cellule en haut à gauche de la grille de la Figure 2 est interprétée par : **si** "le nombre de commandes avec un prix inférieur à 20 est inférieur à 5" **et** "le nombre de commande avec un prix supérieur à 20 est plus grand que 2" **alors** la classe est $C1$.

Étant donné que nous utilisons une représentation équivalente, avec la bonne discrétisation, nous nous attendons à ce que la grille de données optimale correspondante soit capable de détecter le motif contenu dans la variable secondaire. Ainsi, le problème est double : comment trouver la meilleure discrétisation et comment optimiser la grille de données relative. Nous abordons ces deux problèmes simultanément en appliquant une approche par sélection de modèles. Pour ce faire, nous appliquons la méthode MODL (Minimum Optimized Description Length) de Boullé (2006). Le meilleur modèle est choisi en suivant une démarche Maximum A Posteriori (MAP) en maximisant la probabilité $p(\text{Model}|\text{Data})$ du modèle sachant les données. En appliquant la règle de Bayes, cela est équivalent à maximiser $P(\text{Model})p(\text{Data}|\text{Model})$ puisque la probabilité $P(\text{Data})$ est constante quel que soit le modèle. Les modèles considérés prennent en compte la discrétisation de la variable secondaire A et le partitionnement des variables n_k générées. Dans le reste de cette partie, nous décrivons le critère utilisé pour évaluer ces modèles et nous proposons des algorithmes d'optimisation.

2.2 Critère d'Évaluation

Un modèle est entièrement défini par la discrétisation de la variable secondaire (nombre et bornes des intervalles), le partitionnement des variables n_k et la distribution de la variable cible dans chaque cellule de la grille de données qui en résulte. Pour décrire un tel modèle, nous utilisons les notations suivantes.

- N : nombre d'individus (nombre d'enregistrements de la table cible)
- J : nombre de valeurs de la variable à expliquer
- N_s : nombre d'enregistrements de la table secondaire
- K : nombre d'intervalles de discrétisation pour la variable secondaire
- N_k : nombre d'enregistrements de la table secondaire dans le $k^{\text{ème}}$ intervalle ($1 \leq k \leq K$)
- I_k : nombre d'intervalles de discrétisation pour la variable n_k ($1 \leq k \leq K$)
- N_{i_k} : nombre d'individus dans l'intervalle i_k pour la variable n_k ($1 \leq k \leq K$)

– $N_{i_1 i_2 \ldots i_K}$: nombre d'individus dans la cellule $(i_1, i_2, \ldots, i_K)$
– $N_{i_1 i_2 \ldots i_K j}$: nombre d'individus dans dans la cellule $(i_1, i_2, \ldots, i_K)$ pour la valeur cible j

En utilisant les notations ci-dessus, un modèle est complètement défini par les paramètres $\{K, \{N_k\}, \{I_k\}, \{N_{i_k}\}, \{N_{i_1 i_2 \ldots i_K j}\}\}$. Une distribution *a priori* $p(\text{Model})$ est définie sur cet espace de modèles. Elle exploite la hiérarchie naturelle de ces paramètres : le nombre d'intervalles de la variable secondaire A est d'abord choisi, ensuite leurs bornes. Après avoir calculé les variables n_k, une grille multi-variée de dimension K est construite en choisissant pour chaque n_k le nombre d'intervalles, leurs bornes et finalement les effectifs des variables cibles dans chaque cellule. À chaque niveau de cette hiérarchie le choix est supposé être uniforme. Pour le terme de vraisemblance $p(\text{Data}|\text{Model})$, on suppose en outre que les distributions multinomiales des valeurs cibles dans chaque cellule sont indépendantes les unes des autres. En passant au log négatif de $P(\text{Model})p(\text{Data}|\text{Model})$, le critère d'optimisation est donné ci-dessous.

$$\begin{aligned}
&\log N_s + \log \left(\begin{array}{c} N_s + K - 1 \\ K - 1 \end{array} \right) \\
&+ \sum_{k=1}^{K} \log N + \sum_{k=1}^{K} \log \left(\begin{array}{c} N + I_k - 1 \\ I_k - 1 \end{array} \right) \\
&+ \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_K=1}^{I_K} \log \left(\begin{array}{c} N_{i_1 i_2 \ldots i_K} + J - 1 \\ J - 1 \end{array} \right) \\
&+ \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_K=1}^{I_K} \left(\log N_{i_1 i_2 \ldots i_K}! - \sum_{j=1}^{J} \log N_{i_1 i_2 \ldots i_K j}! \right).
\end{aligned} \tag{1}$$

Dans l'équation 1, la première ligne représente le choix de la discrétisation de la variable secondaire : les premier et deuxième termes représentent respectivement le choix du nombre d'intervalles, et les bornes des intervalles. La deuxième et la troisième ligne correspondent au choix de la discrétisation pour chaque variable n_k et les paramètres de la distribution multinomiale pour les valeurs cibles dans chaque cellule de la grille. Le dernier terme représente la probabilité conditionnelle des données sachant le modèle.

Le critère donné par la formule ci-dessus correspond à la probabilité que la grille de données finale (obtenue après la discrétisation de la variable secondaire, tel que décrit auparavant) explique la variable cible connaissant la variable secondaire.

Elle peut également être interprétée comme la capacité d'une grille de données à encoder les classes cibles sachant la variable secondaire, puisque le log négatif de probabilités n'est autre qu'une longueur de codage (Shannon, 1948).

2.3 Algorithme d'Optimisation

Le choix de la discrétisation de la variable secondaire est déterminé par la minimisation du critère vu dans la section 2.2, qui est un problème combinatoire avec 2^{N_s} discrétisations possibles pour la variable secondaire. Par ailleurs, pour chaque discrétisation en K intervalles, il y a $\left(2^N\right)^K$ grilles possibles, qui représentent le nombre des partitions multi-variées des variables $n_1, \ldots, n_K$. Une recherche exhaustive sur tout l'espace des modèles est irréaliste.

Algorithme 1 : Algorithme d'Optimisation

Entrées : $\mathcal{K}$: nombre initial de quantiles,
K_{max} : nombre maximal de discrétisations évaluées
Sortie : D^* : meilleure discrétisation de la variable secondaire,
G^* meilleure grille de données
Pré-condition : $K_{max} \ll \mathcal{K}$

Faire une discrétisation en intervalles de fréquence égale ($\mathcal{K}$ intervalles) ;
Calculer les variables des effectifs initiales ν_k $(\nu_k)_{1 \leq k \leq \mathcal{K}}$;
`/* Initialiser la solution (c*: meilleur coût) */`
$c* \leftarrow \infty$, $D^* \leftarrow$ un intervalle, $G^* \leftarrow$ une cellule;
pour $K \leftarrow 2$ **à** K_{max} **faire**
    $D \leftarrow$ discretisation en K intervalles;
    Estimer le variables effectifs $(n_k)_{1 \leq k \leq K}$ $n_k = \sum_{i=1+\left[\mathcal{K}\frac{k-1}{K}\right]}^{\left[\mathcal{K}\frac{k}{K}\right]} \nu_i$;
    Initialiser G_K (grille de données avec n_k comme variables d'entrée);
    `/* Optimiser la grille de données` G_K `*/`
    $G_K^{'} \leftarrow OptimizeDataGrid(G_K)$;
    si $cost(G_K^{'}) < c^*$ **alors** `// si amélioration du coût`
        `/* retenir la solution améliorée */`
        $c^* \leftarrow cost(G_K^{'})$, $G^* \leftarrow G_K^{'}$, $D^* \leftarrow D$;
    finsi
finpour

L'algorithme 1 fournit une procédure simple pour optimiser la discrétisation de la variable secondaire. La méthode commence par faire une discrétisation fine en intervalles de fréquence égale, ce qui produit $\mathcal{K}$ variables d'effectifs initiales $\nu_1, \ldots, \nu_{\mathcal{K}}$. Ensuite, nous itérons la fusion de ces intervalles initiaux afin de simuler différentes discrétisations en intervalles de fréquence égale. Chaque discrétisation candidate D_k est évaluée en optimisant la grille G_K correspondante. Pour ce faire, nous utilisons l'heuristique d'optimisation des grilles de données multivariées détaillée dans Boullé (2011). A la fin de l'algorithme 1, nous retenons la discrétisation de la variable secondaire avec le coût d'évaluation minimal (voir l'équation 1).

Bien que cet algorithme soit simple et exploite partiellement la richesse des modèles considérés, il demeure une bonne validation de l'approche globale. En priorité pour les travaux futurs, nous prévoyons d'étendre cette procédure d'optimisation afin de mieux explorer l'espace de recherche et de découvrir des modèles de discrétisation plus complexes.

3 Expérimentations

Notre approche a été évaluée par son impact comme une étape de prétraitement à un classifieur de type Bayesien Naïf (BN). Dans ce BN multi-tables, pour une variable numérique secondaire donnée, la grille de données optimale donne une estimation de la densité condition-

	# tables	# variables Sec. num.	# lignes table sec.	# Individus	# valeurs cibles
Mutagenesis-atoms [3]	2	2	1618	188	2
Mutagenesis-bonds [3]	2	4	3995	188	2
Mutagenesis-chains [3]	2	6	5349	188	2
Diterpenses [4]	2	1	30060	1503	23
Miml [5]	2	15	18000	2000	2
Stulong [6]	2	29	10572	1417	2
Xor 2D	2	1	987762	10000	2
Xor 3D	2	1	1843282	10000	2

TAB. 1 – *Description des jeux de données utilisées*

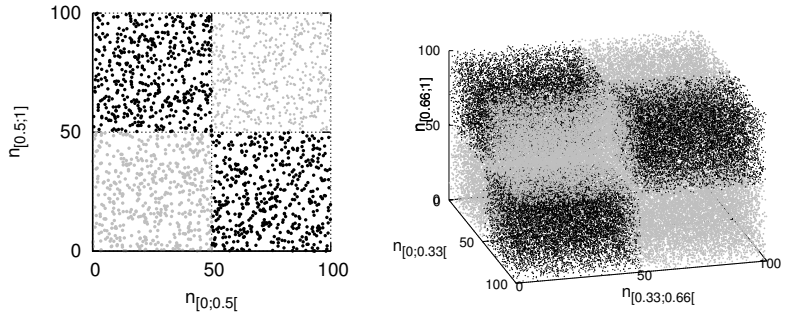
nelle univariée correspondante $P(X_i \mid Y)$. Celle-ci est calculée en considérant les effectifs des valeurs cibles dans chaque cellule. Pour montrer l'apport de notre approche de prétraitement par rapport aux méthodes d'agrégation, pour chaque variable secondaire, la valeur moyenne a été calculée et un BN usuel a été appliqué sur la table à plat résultante. D'autres agrégats ont été testés, à savoir Max, Min et le nombre d'enregistrements dans la table secondaire. Des résultats similaires à ceux décrits ci-dessous ont été obtenus, et ils ne sont rapportés ici en raison du manque d'espace.

Dans nos expérimentations, nous avons considéré différents problèmes de classification basés sur des données synthétiques et réelles. Les caractéristiques de ces données sont indiquées dans la Table 1. Notons que pour chacune des deux bases de données Miml et Stulong, cinq variables peuvent être considérées comme classe cible.

Concernant les données artificielles, la discrétisation idéale est connue à l'avance, et l'étiquette de la variable cible est générée selon une fonction XOR entre les variables d'effectif n_k. La Figure 3 représente les diagrammes de dispersion des données Xor 2D et Xor 3D. Par exemple, dans le modèle Xor 3D (Figure 3(b)), la variable secondaire est discrétisée en trois intervalles : $[0; 0.33[$, $[0.33; 0.66[$ et $[0.66; 1[$. Dans ce schéma plutôt complexe, les points situés, par exemple, près de l'origine (en gris) correspondent à des individus qui ont moins de 50 valeurs dans la table secondaire, respectivement, dans les intervalles $[0; 0.33[$, $[0.33; 0.66[$ et $[0.66; 1[$.

Pour comparer les résultats, nous avons collecté l'aire sous la courbe de ROC (AUC) (Fawcett, 2003) en test en utilisant une validation croisée stratifiée d'ordre 10. Dans toutes les expérimentations, seules les variables secondaires numériques ont été prises en compte et nous avons choisi $\mathcal{K} = 100$ and $K_{max} = 10$ comme paramètres de la procédure d'optimisation (cf. algorithme 1). Évidemment, nous sommes conscients que 10 discrétisations évaluées parmi $o\left(2^{N_s}\right)$ discrétisations candidates ne sont pas suffisantes. Néanmoins, l'objectif de ces expérimentations est principalement d'évaluer le potentiel de l'approche, et de déterminer s'il est utile de travailler sur des procédures d'optimisation plus sophistiquées.

3. http://sourceforge.net/projects/proper/files/datasets/0.1.0/
4. http://cui.unige.ch/~woznica/rel_weka/
5. http://lamda.nju.edu.cn/data_MIMLimage.ashx
6. http://euromise.vse.cz/challenge2003



(a) Motif Xor 2D (b) Motif Xor 3D

FIG. 3 – *Diagrammes de dispersion des bases de données synthétiques. Les couleurs (noir et gris) indiquent l'étiquette de la classe*

La Figure 3 montre les performances de généralisation (AUC en test) obtenues avec un BN en utilisant notre approche de discrétisation (notée MT, pour Multi-Tables) comparées à celles basées sur des variables agrégées. Sur les jeux de données synthétiques (Xor 2D et Xor 3D), notre méthode dépasse largement le BN utilisant la valeur moyenne (notée Avg). Ces résultats ne sont pas surprenants vu la complexité du motif et étant donné que l'agrégation implique une perte d'information. Notre approche est capable de reconnaître le motif contenu dans la variable secondaire et donc de la discrétiser correctement. Ceci est confirmé par la figure 5, qui résume les résultats de la classification obtenue en faisant varier le nombre d'individus dans les jeux de données artificielles. On peut constater qu'avec suffisamment d'individus, notre approche atteint les performances théoriques. Par ailleurs, d'autres expérimentations sur un motif totalement aléatoire montrent que notre méthode est robuste, dans le sens où elle permet de détecter l'absence d'informations prédictives de la variable secondaire (matérialisée par une discrétisation en un seul intervalle et une AUC de l'ordre de 50%).

Sur les données réelles, aucune des deux méthodes ne domine l'autre. En effet, la Figure 3 montre que : (i) notre approche obtient de meilleurs résultats que l'agrégation (mutagenesis-atoms, mutagenesis-chains), (ii) les deux approches obtiennent souvent des résultats similaires (Diterpenses, Miml (mountains, sea, trees), Stulong (CHOLRISK, RARISK)), et dans quelques cas, (iii) l'approche d'agrégation domine notre approche (Miml (sunset), Stulong (OBEZRISK)). Ceci peut être expliqué par le fait que notre critère a besoin d'un grand nombre d'individus pour reconnaître les motifs complexes (ce qui a été montré dans (Boullé, 2011), pour un critère analogue dans le cas d'une seule table), alors que, comme indiqué dans la Table 1, les bases de données réelles utilisées sont de relativement petites tailles. Par ailleurs, nous rappelons que l'algorithme 1 est assez simple et n'exploite pas tout le potentiel du critère de discrétisation, en se limitant aux intervalles de fréquence égale.

Nous tenons à souligner que, bien que les résultats de notre approche ne soient pas toujours meilleurs que pour une méthode par agrégation, ceci pourrait être expliqué par l'exploration insuffisante de l'espace des modèles. Par ailleurs, l'approche est capable de détecter des motifs complexes (cf. Figure 3) et fournir des règles interprétables pour l'utilisateur, ce qui est difficilement possible pour les méthodes par agrégation.

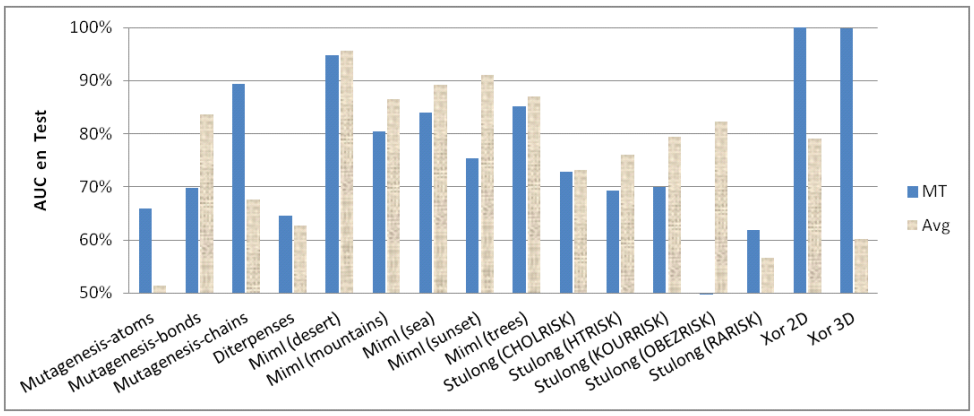
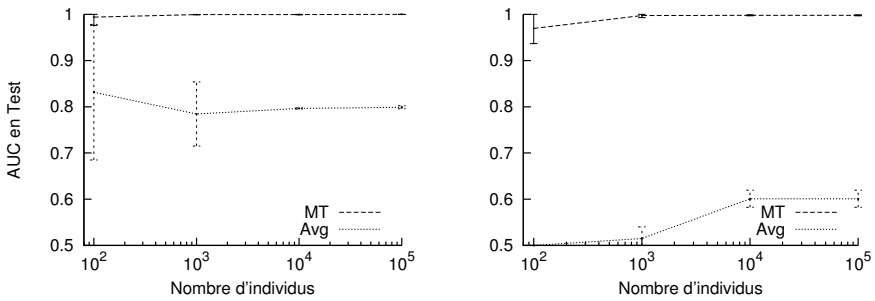


FIG. 4 – *Résultats obtenus sur les données artificielles et réelles*



(a) Motif Xor 2D (b) Motif Xor 3D

FIG. 5 – *Résultats obtenus sur les données artificielles*

Afin de montrer comment interpréter la discrétisation générée, prenons par exemple, la base de données Stulong (composée d'une table cible Patient en relation un-à-plusieurs avec une table Examen). On s'intéresse en particulier à la variable cible CHOLRISK (trois valeurs cibles : Normal, Risquée et manquant) qui indique si le patient présente un risque de cholestérol élevé, et la variable secondaire CHLSTMG (dans la table Examen) qui décrit pour chaque examen le taux de cholestérol. L'algorithme 1 discrétise la variable CHLSTMG en deux intervalles, à savoir $]inf, 228.5[$ et $[228.5, sup[$. La Table 2 représente la grille de données optimale correspondant à cette discrétisation (les valeurs entre parenthèses sont les effectifs des valeurs cibles dans chaque cellule). Ce tableau peut être interprété comme un ensemble de quatre règles de classification, un pour chaque cellule.

Par exemple la cellule en haut à gauche est équivalente à la règle : **Si** au moins quatre examens ont un taux de cholestérol inférieur à 228.5 mg **et** aucun examen avec un taux de cholestérol supérieur à 228.5 mg, **alors** la classe est **Normal** (ce qui signifie qu'il n'y a pas de risque de cholestérol).

$[4; sup[$	$(100; 0; 0)$	$(77; 23; 0)$
$[0; 3]$	$(81; 18; 1)$	$(40; 58; 2)$
$n_{]inf,228.5[} \times n_{[228.5,sup[}$	0	$[1, sup[$

TAB. 2 – *Table de contingence relative à la discrétisation de la variable CHLSTMG*

4 Conclusion

Dans cet article, nous avons présenté une nouvelle approche pour discrétiser les variables numériques secondaires dans le cadre de la fouille de données multi-tables. La méthode consiste à projeter les variables secondaires numériques vers la table cible. Ceci est réalisé en discrétisant ces variables, puis en calculant pour chaque individu le nombre d'enregistrements de la table secondaire dans chaque intervalle. En outre, nous avons proposé un critère qui évalue dans quelle mesure une discrétisation préserve la corrélation avec la variable cible. Nous avons également décrit un algorithme d'optimisation pour trouver une estimation de la meilleure discrétisation en intervalles de fréquence égale. Cependant, cette procédure n'exploite pas tout le potentiel du critère. Comme travaux futurs, nous envisageons d'étendre notre algorithme afin de mieux explorer l'espace de recherche et de découvrir des motifs de discrétisation plus complexes.

Cette étude nous a montré, à travers des expérimentations sur des jeux de données artificielles, que le critère d'évaluation ainsi que la procédure de discrétisation permettent de découvrir des variables secondaires pertinentes et assurer des taux de prédiction importants. Toutefois, dans le cas des données réelles, nous avons besoin de chercher des jeux de données de plus grande taille, afin de mieux évaluer notre approche et de la comparer à d'autres techniques de fouille de données multi-tables.

Références

Alfred, R. (2009). Discretization Numerical Data for Relational Data with One-to-Many Relations. *Journal of Computer Science 5*(7), 519–528.

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. (2011). Data Grid Models for Preparation and Modeling in Supervised Learning. In I. Guyon, G. Cawley, G. Dror, et A. Saffari (Eds.), *Hand on pattern recognition : Challenges in Machine Learning*, pp. 99–130. Microtome Publishing.

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

Fawcett, T. (2003). ROC graphs : Notes and practical considerations for researchers. Technical report, Technical Report HPL-2003-4, Hewlett Packard Laboratories.

Fayyad, U. M. et K. B. Irani (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Thirteenth International Joint Conference on Artificial Intelligence*, pp. 1022–1027.

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

Knobbe, A. J. (2004). *Multi-Relational Data Mining*. Ph. D. thesis, Faculteit Wiskunde en Informatica, Universiteit Utrecht.

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

Knobbe, A. J. et E. Ho (2005). Numbers in multi-relational data mining. *Lecture notes in computer science 3721*, 544.

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. New York, NY, USA : Springer-Verlag.

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

Van Laer, W., L. De Raedt, et S. Džeroski (1997). On multi-class problems and discretization in inductive logic programming. In Z. W. Ras et A. Skowron (Eds.), *Proceeding of the 10th International Symposium on Foundations of Intelligent Systems, ISMIS '97*, Charlotte, North Carolina, USA, pp. 277–286. Springer-Verlag.

Summary

In Multi-Relational Data Mining (MRDM), data are 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. Variable pre-processing (including discretization and feature selection) within this multiple table setting differs from the attribute-value case. Besides the target variable information, one should take into account the relational structure of the database. In this paper, we focus on numerical variables located in a non target table. We propose a criterion that evaluates a given discretization of such variables. The idea is to summarize for each individual the information contained in the secondary variable by a feature tuple (one feature per interval of the considered discretization). Each feature represents the number of values of the secondary variable ranging in the corresponding interval. These count features are jointly partitioned by means of data grid models in order to obtain the best separation of the class values. We describe a simple optimization algorithm to find the best equal frequency discretization with respect to the proposed criterion. Experiments on a real and artificial data sets reveal that the discretization approach helps one to discover relevant secondary variables.