An Evaluation Criterion for Itemset Based Variable Construction in Multi-Relational Supervised Learning

Dhafer Lahbib[1,2], Marc Boullé[1], and Dominique Laurent[2]

[1] Orange Labs - 2, avenue Pierre Marzin, 23300 Lannion
{dhafer.lahbib,marc.boulle}@orange.com
[2] ETIS-CNRS-Universite de Cergy Pontoise-ENSEA, 95000 Cergy Pontoise
{dominique.laurent}@u-cergy.fr

Abstract. 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 in one-to-many relationship. In this paper, we suggest an itemset based framework for constructing variables from secondary tables and evaluate their conditional information for the supervised classification task. We introduce a space of itemset based models in the secondary table and 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.

Keywords: Supervised Learning, Multi-Relational Data Mining, one-to-many relationship, variable selection, variable construction

1 Introduction

Learning from relational data has recently received increasing attention in the literature. Multi-Relational Data Mining (MRDM) was introduced in [6] to address knowledge discovery techniques from multiple relational tables. The common point between MRDM techniques is that they need to transform the relational representation. In Inductive Logic Programming ILP [3], data is recoded as logic formulas. Other methods, known as propositionalisation [7], try to flatten the relational data by creating new variables. To the best of our knowledge, few studies have treated the variable preprocessing problem in the MRDM context with one-to-many relationship. Some works in ILP operate by selecting predicates in order to reduce the search space during the learning step [1, 4].

The purpose of this paper is to construct new variables from a secondary table having a one-to-many relation with the target table and to evaluate their relevance for the task of supervised classification. In order to take into account the risk of overfitting, which dramatically increases with the number of potential constructed variables, we introduce a space of itemset based models in the secondary table and conditional density estimation of the resulting constructed

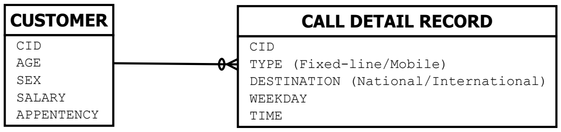


Fig. 1. Relational schema of a CRM database

variable. Then a prior distribution is defined on this model space. As a result, we obtain a parameter-free relevance criterion for the constructed variables. We illustrate our approach through the following example.

Example 1. Figure 1 shows an extract of a Customer Relationship Management (CRM) relational database schema. The table *Customer* is the target table, where each customer (or individual) data is stored in a single row. On the other hand, *Call detail record* (CDR) is a secondary table linked to *Customer* through the foreign key CID. Thus, in the CDR table, several rows are related to a single customer from the *Customer* table. In our example, the problem is to identify the customers likely to be interested in a particular product. This problem turns into a classification problem where the target variable is the boolean variable *Appetency*, which denotes whether the customer is likely to order that product.

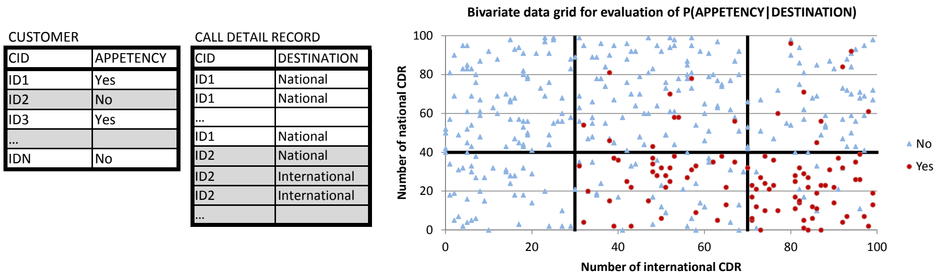
To do so, we consider itemsets defined as a conjunction of expressions of the form $(x \in S_x)$ where x is a variable of the CDR table and S_x is either an interval (if x is a numerical variable) or a set of values (if x is a categorical variable). Assuming that $WeekDay$ and $Destination$ are categorical variables and $Time$ is a numerical variable, π: $(WeekDay \in \{Saturday, Sunday\}) \wedge (Time \in (10{:}00{:}00\,; 11{:}30{:}00]) \wedge (Destination \in \{International\})$ is an itemset. This itemset π allows constructing a new binary variable A_π, according to whether the secondary records are covered or not by the itemset.

Our relevance criterion is the sum of two criteria: (i) an evaluation criterion assessing the relevance of A_π w.r.t. the target variable (as defined in [8]), and (ii) a construction criterion assessing the encoding cost of the itemset π.

The rest of this paper is organized as follows. Section 2 recalls the method [8] dealing with the case of a binary secondary variable. Section 3 introduces the space of constructed itemset based secondary variables and presents their evaluation criterion. Section 4 gives a summary and discusses future work.

2 Evaluation of Binary Secondary Variables

The method introduced in [8] is able to evaluate the relevance of a binary secondary variable A with values a and b. In this case, each individual is described by a bag of secondary values a and b, and summarized without loss of information by the numbers n_a of a and n_b of b in this bag. Thus, the whole information about A can be captured by considering jointly the pair (n_a, n_b) of primary variables. We emphasize that n_a and n_b are considered jointly so as to preserve information, as illustrated in Figure 2 in the context of Example 1.



CUSTOMER

CID	APPETENCY
ID1	Yes
ID2	No
ID3	Yes
...	
IDN	No

CALL DETAIL RECORD

CID	DESTINATION
ID1	National
ID1	National
...	
ID1	National
ID2	National
ID2	International
ID2	International
...	

Fig. 2. Evaluation of the secondary variable *Destination* for the prediction of the target variable *Appetency*. Here, customers with a small number of national CDR and a large number of international CDR are likely to have the value *Yes* of *Appetency*.

In this setting, $P(Y \mid A)$ is equivalent to $P(Y \mid n_a, n_b)$. Bivariate data grid models [2] are used to qualify the information contained in the variable pair (n_a, n_b). The values of n_a and n_b are partitioned jointly into intervals, thus giving a partitioning of the data into a grid whose cells are defined by intervals pairs. The target variable distribution is defined locally in each cell. Therefore, the purpose is to find the optimal bivariate discretization which maximizes the class distribution, in other words, we look for the optimal grid with homogeneous cells according to the class values. To do so, we introduce the following notation:

- N : number of individuals (number of target table records)
- J : number of target values (classes),
- I_a,I_b : number of discretization intervals respectively for n_a and n_b
- $N_{i_a..}$: number of individuals in the interval i_a $(1 \leq i_a \leq I_a)$ for variable n_a
- $N_{.i_b.}$: number of individuals in the interval i_b $(1 \leq i_b \leq I_b)$ for variable n_b
- $N_{i_a i_b.}$: number of individuals in the cell (i_a, i_b)
- $N_{i_a i_b j}$: number of individuals in the cell (i_a, i_b) for the target value j

Applying a Bayesian model selection approach, our evaluation criterion $c_e(A)$ to assess the relevance of a secondary binary variable is defined as follows.

$$c_e(A) = \log N + \log N + \log \binom{N + I_a - 1}{I_a - 1} + \log \binom{N + I_b - 1}{I_b - 1} \quad (1)$$
$$+ \sum_{i_a=1}^{I_a} \sum_{i_b=1}^{I_b} \log \binom{N_{i_a i_b.} + J - 1}{J - 1} + \sum_{i_a=1}^{I_a} \sum_{i_b=1}^{I_b} \log \frac{N_{i_a i_b.}!}{N_{i_a i_b 1}! N_{i_a i_b 2}! \dots N_{i_a i_b J}!}$$

Details about this criterion and the optimization algorithm can be found in [2]. Beyond the evaluation of binary secondary variables, our goal is to extend the method to numerical and categorical secondary variables to capture the potential correlations that may exist between them. In the next section, we introduce a similar criterion for itemset-based models defined over secondary variables, in order to take into account this multivariate correlation and to benefit from the potential of itemset-based classification rules [5].

3 Itemset Based Variable Construction

Itemset Based Construction Model. Based on the classification model of [5], an itemset is a conjunction of terms of the form $(x \in S_x)$ where x is a variable from the secondary table, and S_x is either an interval if x is numerical, or a set of values if x is categorical.

Every itemset π is associated with a boolean variable A_π, which is true for secondary records that are covered by π, and false otherwise. A_π is called an *Itemset Based Construction Model variable*, or IBCM variable, for short.

Our working model space is thus the space of all itemsets. To apply the Bayesian approach, we first define a prior distribution on the set of all possible itemsets. To this end, we introduce the following notation, where for every secondary variable x, $dom(x)$ denotes the set of all possible values over x.

- N_s : number of records in the secondary table
- m : number of (categorical or numerical) variables in the secondary table
- $X = \{x_1, \ldots, x_k\}$: set of k variables occurring in the itemset $(k \leq m)$
- X_{cat} (resp. X_{num}) : set of categorical (resp. numerical) variables occurring in the itemset $(X = X_{cat} \cup X_{num})$
- V_x : number of values of a categorical variable x $(V_x = |dom(x)|)$
- I_x : number of intervals or groups of a variable x
- $\{i(v_x)\}_{v_x \in dom(x)}$: set of indexes of groups to which v_x is affected (one index per value, either 1 or 2 for inside or outside of the itemset)
- $\{N_{i(x).}\}_{1 \leq i \leq I_x}$: number of records in interval i_x of numerical variable x
- $i_{x_1}, \ldots, i_{x_k}$: indexes of groups of categorical variables (or intervals of numerical variables) occurring in the itemset

MODL hierarchical prior. We use the following distribution prior on itemsets, called the MODL hierarchical prior. Notice that a uniform distribution is used at each stage[3] of the parameters hierarchy of the models:

1. the number of variables k in the itemset is uniformly distributed in $[0; m]$
2. for a number k of variables, every set of k constituent variables of the itemset is equiprobable (given a drawing with replacement)
3. for a categorical variable x in the itemset, the number of groups is 2 $(I_x = 2)$
4. for a numerical variable x in the itemset, the number of intervals I_x is either 2 or 3 (with equiprobability)
5. for a categorical (or numerical) variable x and a number of groups (or intervals), every partition of x into I_x groups (or intervals) is equiprobable (cf. $\{i(v_x)\}_{v_x \in dom(x)}$ for groups and $\{N_{i(x).}\}_{1 \leq i \leq I_x}$ for intervals)
6. for a categorical variable x, for a value group i_x of this variable, belonging to the itemset or not is equiprobable
7. for a numerical variable x with 2 intervals, for an interval i_x of this variable, belonging to the itemset or not is equiprobable. In the case of 3 intervals, the itemset interval is necessarily the middle one.

[3] It does not mean that the hierarchical prior is a uniform prior over the itemset space, which would be equivalent to a maximum likelihood approach.

Using the definition of the model space and its prior distribution, we now express the prior probabilities of our Itemset Based Variable Construction model.

Construction cost of an IBCM variable. The construction cost $c_c(A_\pi)$ of an IBCM variable A_π associated with an itemset π is defined as follows:

$$\begin{aligned} c_c(A_\pi) = {} & \log(m+1) + \log \binom{m+k-1}{k} \\ & + \sum_{x \in X_{cat}} \log \mathcal{S}(V_x, 2) + \log 2 \\ & + \sum_{x \in X_{num}} \log 2 + \log \binom{N_s - 1}{I_x - 1} + \log(1 + \mathbb{1}_{\{2\}}(I_x)) \end{aligned} \tag{2}$$

It can be seen from Formula 2 that the cost of an IBCM variable is the negative logarithm of probabilities, thus expressing a coding length according to Shannon [9]. Here, $c_c(A_\pi)$ may be interpreted as a variable construction cost, that is the coding cost of the itemset π.

In Formula 2, the first line stands for the choice of the number of variables occurring in the itemset and for the choice of these variables among all variables in the secondary table. The second line is related to the choice of the groups and the values involved in the itemset for categorical variables, where the number of partitions of V_x values into 2 groups is $\mathcal{S}(V_x, 2)$ ($\mathcal{S}$ stands for Stirling number of the second kind). The third line deals with numerical variables, *i.e.,* the choice of the number of intervals, their bounds and the one involved in the itemset ($\mathbb{1}_{\{2\}}$ is the characteristic function of the set $\{2\}$, that is, $\mathbb{1}_{\{2\}}(I_x) = 1$ if $I_x = 2$ and $\mathbb{1}_{\{2\}}(I_x) = 0$ otherwise).

The new variable A_π that we have built is no other than a binary secondary variable which can be evaluated using the cost $c_e(A_\pi)$ of Formula 1. Thus, in order to evaluate the overall relevance $c_r(A_\pi)$ of A_π, we have to take into account its construction cost $c_c(A_\pi)$ as well as the evaluation cost $c_e(A_\pi)$:

$$c_r(A_\pi) = c_c(A_\pi) + c_e(A_\pi) \tag{3}$$

The construction cost $c_c(A_\pi)$ acts as a regularization term. Constructed variables based on complex itemsets, with multiple constituent variables in the itemset, are penalized compared to simple constructed variables. Let $\pi_\emptyset$ be the empty itemset, with no constituent variable, where no secondary record is covered by the itemset. The relevance cost of the empty itemset constructed variable is

$$c_r(A_{\pi_\emptyset}) = \log(m+1) + 2 \log N + \log \frac{N!}{N_1! N_2! \dots N_J!} \tag{4}$$

$$= N.Ent(Y) + O(\log N) \tag{5}$$

where $Ent(Y)$ is the Shannon entropy of the target variable Y, and N_j $(1 \leq j \leq J)$ is the number of individuals associated with class number j over Y.

Therefore, any itemset π whose IBCM variable A_π has a relevance cost beyond the relevance cost of the empty itemset constructed variable can be discarded, as being less informative than the target variable alone.

4 Conclusion

In this paper, we have proposed an approach for constructing new variables and assessing their relevance in the context of multi-relational supervised learning. The method consists in defining an itemset in a secondary table, leading to a new secondary variable that collects whether secondary records are covered or not by the itemset. The relevance of this new variable is evaluated using a bivariate supervised data grid model [8], which provides a regularized estimator of the conditional probability of the target variable. To avoid overfitting, we applied a Bayesian model selection approach for the itemset based construction model and the conditional density evaluation model. In doing so, we obtained an exact criterion for the posterior probability of any constructed variable.

Our future work will aim at providing search heuristics to explore the space of constructed variables and keep the most relevant ones with their estimated conditional probability. Classifiers using a univariate preprocessing like Naive Bayes or Decision Trees could then be extended to multi-relational data.

References

1. Alphonse, E., Matwin, S.: Filtering multi-instance problems to reduce dimensionality in relational learning. Journal of Intelligent Inf. Syst. 22(1), 23–40 (2004)
2. Boullé, M.: Optimum simultaneous discretization with data grid models in supervised classification A Bayesian model selection approach. Advances in Data Analysis and Classification 3(1), 39–61 (2009)
3. Džeroski, S., Lavrač, N.: Relational Data Mining. Springer-Verlag (2001)
4. Fürnkranz, J.: Dimensionality reduction in ILP: A call to arms. In: Proc. of the IJCAI-97 Workshop on Frontiers of Inductive Logic Programming. pp. 81–86 (1997)
5. Gay, D., Boullé, M.: A bayesian approach for classiffication rule mining in quantitative databases. In: ECML/PKDD'2012. Springer Verlag (2012), to appear
6. Knobbe, A.J., Blockeel, H., Siebes, A., Van Der Wallen, D.: Multi-Relational Data Mining. In: Proc. of Benelearn '99 (1999)
7. Kramer, S., Flach, P.A., Lavrač, N.: Propositionalization approaches to relational data mining. In: Relational data mining, pp. 262–286. Springer-Verlag (2001)
8. Lahbib, D., Boullé, M., Laurent, D.: Informative variables selection for multi-relational supervised learning. In: Proc. of the 7th International Conference on Machine Learning and Data Mining in Pattern Recognition. pp. 75–87 (2011)
9. Shannon, C.: A mathematical theory of communication. Technical report. Bell systems technical journal (1948)