Informative Variables Selection for Multi-Relational Supervised Learning

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

[1] France Telecom R&D - 2, avenue Pierre Marzin, 23300 Lannion
{dhafer.lahbib,marc.boulle}@orange-ftgroup.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. To cope with this one-to-many setting, most of the existing approaches try to transform the multi-table 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 estimate 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 the approach allows to potentially identify relevant one-to-many variables. In this article, we focus on binary variables because of space constraints.

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

1 Introduction

Most of existing data mining algorithms are based on an attribute-value representation. In this flat format, each record represents an individual and the columns represent variables describing these individuals. In real life applications, data usually present an intrinsic structure which is hard to express in a tabular form. This structure may be naturally described using the relational formalism where objects are distributed on several tables. That's why learning from relational data have recently received a lot of attention in the literature. The term Multi-Relational Data mining was initially introduced by [10] to describe a new family of knowledge discovery techniques from multi table relational data. The common point between these techniques is that they need to transform the relational representation : in Inductive Logic Programming ILP [5], data is recoded as logic

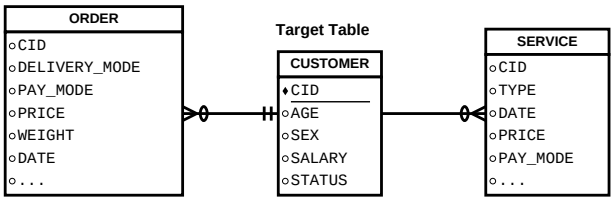


Fig. 1: Relationnel schema of a CRM database

formulas. This causes scalability problems especially with large-scale data. Other methods called by Propositionalisation [11] try to flatten the relational data by creating new variables. These variables aggregate the information contained in non target tables in order to obtain a classical attribute-value format. Consequently, not only we lose the naturally compact initial representation but there is a risk of introducing statistical bias because of potential dependencies between the newly added variables.

Let us take as an example, a Customer Relationship Management (CRM) problem. Figure 1 represents an extract of a virtual CRM relational database schema. The problem may be, for instance, to identify the customers likely to be interested in a certain product or service, which turns into a customer classification problem. For this reason, we focus our analysis on the customers in the target table. The target variable is the *Status* attribute which denotes whether the customer has already ordered a particular product. The *Customer* table is related to non target tables, *Order* and *Service*, with a one-to-many relationship. Predicting whether the customer would be interested in a product does not depend only on the information of that customer. Those describing the other products that he has ordered may potentially be very informative. Variables like the product *Weight* or *Price* may present correlations with the target variable and may be very useful to predict its value.

Studying the predictive importance of variables located in non target tables raises several difficulties because of one-to-many relationships with the target table. In the attribute-value mono table case, each individual has a single value per variable. While in multi table setting, for a non target table variable, an individual may have a value list (eventually empty) of varying size. To the best of our knowledge, only few studies in the literature 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 large search space during the learning step [7, 1]. Others cope directly with the initial representation : in [9] the authors operate simply by propagating the target variable toward non target tables. Then they compute the Information Gain in the same way as in the monotable case. By doing such a propagation, there is a severe risk of overfitting certain individuals : those having more related records in the non target table will have more importance.

The purpose of this article is to study the relevance of a secondary variable situated in a non target table having a one-to-many relation with the target one[3]. This relevance is evaluated by estimating the conditional probability $P(Y \mid A)$, where Y is the target variable and A is a secondary variable. This univariate preprocessing extended to the relational context is of a great interest for filter feature selection [8] or as preprocessing step for classifiers such as Naive Bayes or Decision Tree.

The remainder of this paper is organized as follows. Section 2 summarizes our approach in the case of a binary secondary variable. In Section 3 we evaluate the approach on artificial and real datasets. Finally, Section 4 gives a summary and discusses future work.

2 Approach Illustration

Let us remember that our goal is to evaluate the relevance of a variable located in a non target table. To simplify the problem, let us take the simplest case: a binary variable with two values a and b. In this case, each individual is described by a bag of values among a and b [4]. Given an individual, all that we need to know about the secondary variable are the number of a and the number of b in the bag related to that individual (we denote them respectively n_a and n_b). Thus, the whole information about the initial variable can be captured by considering jointly the pair (n_a, n_b). We emphasize that the two variables must be considered jointly so that we preserve the information about the proportions ($\frac{n_a}{n_a+n_b}$ and $\frac{n_b}{n_a+n_b}$) and the cardinalities ($n_a + n_b$) of the initial values in each bag.

By doing so, the conditional probability $P(Y \mid A)$ is equivalent to $P(Y \mid n_a, n_b)$. To qualify the information contained in the variable pair (n_a, n_b), we suggest to use bivariate grid models [4]. The idea is to jointly discretize the two numeric variables into intervals. Individuals are then partitioned into a data 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, obtain the optimal grid with homogeneous cells according to the class values (cf. figure 2). It is an interpretable representation since it shows the distribution of the individuals on the data grid while jointly varying the two variables:

- by moving on both axis, we vary the numbers of a and b,
- on the first diagonal (including the origin point), we vary the total number $n_a + n_b$,
- on the opposite diagonal, we vary the ratios $\frac{n_a}{n_a+n_b}$ and $\frac{n_b}{n_a+n_b}$,

[3] The one-to-one relationship is equivalent to the monotable case. For simplification reasons, we limit the relationship to the first level : tables directly related to the target one.

[4] This is different from the attribute-value setting, where for a given variable, an individual can only have a single value.

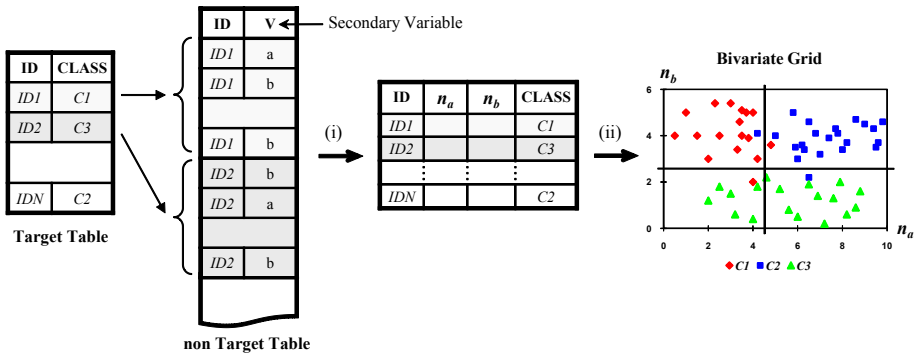


Fig. 2: Illustration of the Approach

The optimal bivariate grid is obtained through a greedy optimization procedure by varying at each step the interval bounds. This procedure is guided using an MDL criterion which evaluates a given bivariate discretization. We follow the MODL (Minimum Optimized Description Length) approach [4] to define this criterion in the next section.

2.1 Evaluation criterion

In the MODL approach, the joint partitioning of two continuous variables is transposed into a model selection problem. The best model is chosen according to a maximum a posteriori (MAP) approach: maximizing the probability p(Model|Data) of the model given the data. By applying the Bayes rule, this is equivalent to maximizing P(Model)p(Data|Model) since the probability P(Data) is constant under varying the model.

Notation 1.

- N *: number of individuals (number of target table records)*
- J *: number of target values,*
- 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

The partitioning model parameters are the numbers of intervals I_a and I_b, the frequencies of the intervals $\{N_{i_a..}\}$ and $\{N_{.i_b.}\}$ and the distribution of the target values $\{N_{i_a i_b j}\}$ in each cell (i_a, i_b). A bivariate discretisation model is completely defined by the parameters $\{I_a, I_b, \{N_{i_a..}\}, \{N_{.i_b.}\}, \{N_{i_a i_b j}\}\}$. The prior distribution p(Model) is defined on this model space. It exploits the natural hierarchy of the parameters: the number of intervals are first chosen, then the bounds of the intervals and finally the frequencies of the target values in each cell. At each stage of this hierarchy the choice is supposed to be uniform.

For the likelihood term $p(\text{Data}|\text{Model})$, we assume further that the multinomial distributions of the target values in each cell are independent from each other. By applying the Bayes rule we can calculate the exact prior probabilities $p(\text{Model})$ and the likelihood $p(\text{Data}|\text{Model})$. Taking the negative log of these probabilities, the supervised bivariate discretization criterion of two continuous variables n_a et n_b is provided in equation 1.

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

The first five terms stand for the prior probability: choosing the numbers of intervals, their frequencies and the distribution parameters for the target values in each grid cell. The last term represents the conditional likelihood of the data given the model.

2.2 Grid Optimisation

The bivariate discretization criterion is optimized starting from an initial random solution and alternating partial optimization per variable. Bottom-up greedy heuristics are used for theses partial optimizations. Pre and post optimization steps are used to escape from eventual local optima. The overall complexity of the algorithm is $O\left(JN^{\frac{3}{2}}\log(N)\right)$ [4].

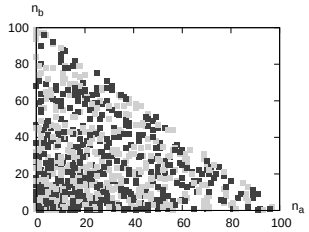
3 Experiments

This section evaluates our approach in the case of a binary secondary variable. We use real and artificial datasets. Using artificial datasets allows controlling the relationship between the secondary variable and the target.

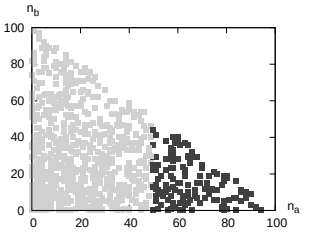
3.1 Protocol

The intrinsic predictive power of a binary secondary variable is evaluated using a univariate classifier based on this variable only. For each secondary variable, we apply the approach described in section 2 in order to obtain the corresponding optimal bivariate data grid. The data grid built during the training step is used as a decision table. To classify an individual in the test set, we place it in the data grid. The predicted target value is the most frequent class in the corresponding cell (according to the collected frequencies during training). We call this classifier, henceforth, the Bivariate Grid (BG) classifier.

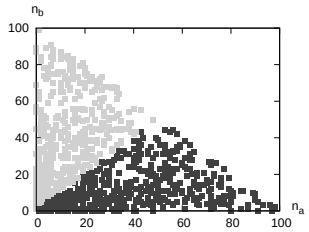
The relevance of a secondary variable is evaluated using the accuracy rate and Area Under ROC Curve AUC [6] of the corresponding classifier. We also report the optimal grid size which provides some insight related to the underlying pattern complexity. For all experiments, we use a 10 fold stratified cross validation.



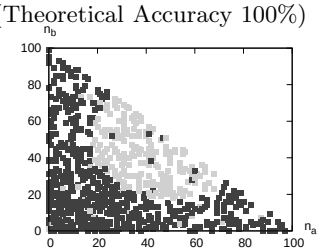
(a) Random pattern
$P(c = 0) = P(c = 1) = 0.5$
(Theoretical Accuracy 50%)



(b) Cardinality based pattern
if $n_a \geq T$ then $c = 1$
else $c = 0$
(Theoretical Accuracy 100%)

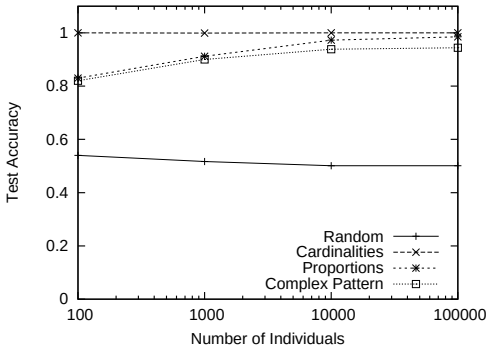


(c) Proportion based pattern
if $\frac{n_a}{n_a+n_b} \geq R$ then $c = 1$
else $c = 0$
(Theoretical Accuracy 100%)

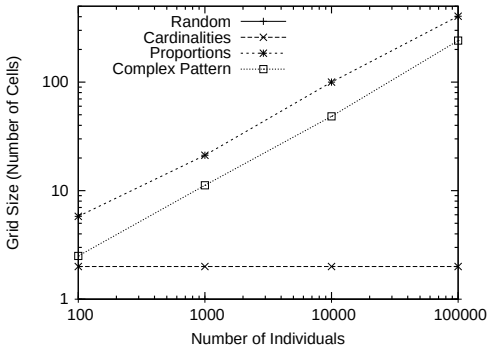


(d) "Complex" pattern
if $(n_a - M_1)^2 + (n_b - M_2)^2 \geq R$
then $c = 1$ with 5% of noise
else $c = 0$ with 5% of noise
(Theoretical Accuracy 95%)

Fig. 3: Scatterplots of the artificial datasets (T, R, M_1 and M_2 are thresholds)



(a) Test Accuracy



(b) Grid size

Fig. 4: Results obtained on artificial datasets

3.2 Artificial datasets

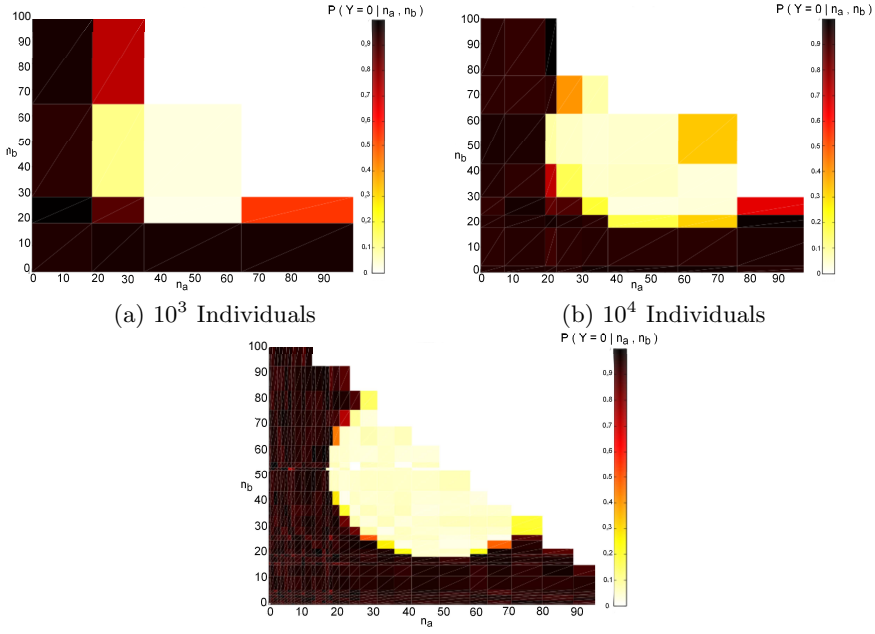
The artificial data sets have a relational structure consisting of two tables: a target table in one-to-many relation with a secondary table which contains only one binary variable. The data generator takes as parameters the number of individuals (target table records) and the maximum number of records related to each individual in the secondary table. The non target records are uniformly distributed between the 'a' and 'b' values; and the target value (binary) is generated according to a controlled pattern in the secondary variable.

Figure 3 depicts the scatter plots of the generated datasets as well as the subsequent patterns (for each individual the number of records is uniformly chosen in [0..100]). The random pattern where there is no dependency between the secondary variable and the target one allows us to evaluate the robustness of the approach. The other datasets aim to check whether the subsequent pattern, more or less complex, could easily be found.

Figure 4a illustrates the classification results for the 5 artificial datasets for different individuals number. At first, the results show that the method can easily detect a totally random pattern. The absence of predictive information in the data grid is materialized by a bivariate discretization in one single cell (Figure 4b) and a test accuracy of about 50%. The method can also detect complex patterns. Figure 4a shows that the classification performance improves with the number of individuals in the database. With enough individuals, the accuracy reaches approximately the theoretical performance. The grid size varies according to the complexity of the bivariate pattern: for a fairly simple pattern based on cardinalities, the grid is always composed of two cells; for more complex patterns, the grid size increases with the individuals number, thereby approximating finely the pattern.

To see some examples of the optimal data grids, we show in Figure 5 the obtained bivariate partitioning on the dataset of Figure 3d with respectively 10^3, 10^4 and 10^5 individuals. The figures represent the probability distribution of the target variable $P(Y|n_a, n_b)$ in each cell. The resulting grids show that the more we have individuals in the dataset, the more our approach is able to recognize the pattern.

In order to study the advantage of our approach over aggregation based methods, we compare the performance of our approach with attribute-value classifiers using aggregated features. We introduce two classifiers: a Naive Bayes (NB) [12] and Best Univariate (BU). Both of these classifiers are based on a univariate preprocessing: first the optimal discretization model for each input feature is found (we use the MODL method applied to the univariate case [2], i.e. by considering each feature independently from the others). After that, for the NB classifier, the univariate conditional probabilities are deduced from the target values frequencies in the subsequent intervals. The BU classifier looks for the best feature which maximizes the probability that its discretization model explains the feature. In order to classify an instance in test, the corresponding interval is determined with respect to the trained discretization model. The predicted target value is then the most frequent one in that interval.



(a) 10^3 Individuals (b) 10^4 Individuals

(c) 10^5 Individuals

Fig. 5: Examples of Bivariate Grids obtained with the dataset of Figure 3d for different numbers of individuals

We evaluate different variants of these classifiers according to the used aggregation feature set. Experimentations are performed on the datasets of Figure 3. To summarize, the evaluated classifiers are:

- Bivariate Grid: the bivariate optimum data grid obtained as described in our approach by considering jointly the pair (n_a, n_b)
- Best Univariate the BU classifier using all the aggregation features: n_a, n_b, $n_a + n_b$, $\frac{n_a}{n_a+n_b}$ and $\frac{n_b}{n_a+n_b}$
- Naive Bayes (A): the NB classifier using all the previous features
- Naive Bayes (C): the NB classifier using cardinality based aggregation features n_a and n_b
- Naive Bayes (R): the NB classifier using ratios $\frac{n_a}{n_a+n_b}$ and $\frac{n_b}{n_a+n_b}$.

Figure 6 depicts the test accuracy obtained with these classifiers by varying the number of individuals. There are several observations that can be made from looking at these histograms. First, for datasets with a cardinality based pattern (Figure 6a), all classifiers do well (since they are using n_a) except NB(R) which is based only on ratios $\frac{n_a}{n_a+n_b}$ and $\frac{n_b}{n_a+n_b}$.

Next, for the proportion based pattern (Figure 6b), as one could expect, all classifiers exploiting the feature $\frac{n_a}{n_a+n_b}$ (NB(A), NB(R) and BU) recognize easily

the pattern. Interestingly, the Bivariate Grid classifier even if it exploits only cardinality features, it is able with enough individuals to approximate the pattern. In contrast, NB(C) which is using cardinality features has less performances.

In Figure 6c, the pattern is harder since it implies more than one feature. Not surprisingly, the BG classifier carries the best accuracies and reaches the theoretical performances. BU is unable to recognize the pattern as it is a univariate classifier which is not sufficient. What is a bit surprising is that the performance of the Naive Bayes classifier (NB(A) and NB(C)) is pretty good (although less than that of BG). The reason for that is that the univariate preprocessing carried out by the NB is fairly good and that the NB classifier efficiently exploits all the agregated features.

To summarize, in comparing the results obtained when using BG versus NB, both classifiers have almost the same performance on relatively simple patterns. However, with a more complex pattern (Figure 6c) and with enough individuals, the former has better accuracy. We believe that this is because NB performs a univariate preprocessing, whereas, BG makes a bivariate one. The same observation can be made by comparing BG to BU. Looking further at the used aggregation features, it can be seen that by considering only two features (n_a and n_b), our approach performs as well as NB and BU classifiers, and even better, than when using more features. As expected, the pair (n_a, n_b) preserves the whole information contained in the initial non target table variable, namely the information on cardinalities and ratios.

3.3 Stulong dataset

STULONG [13] is a medical dataset collected during a 20-year longitudinal study on risk factors of atherosclerosis [5] in a population of 1 417 middle aged men. In this dataset, we are interested in tables *Entry* and *Control*. In table *Entry*, the target table, the variables describe the patients' properties. Among these variables only some of them can be considered as target ones. The others do not concern us because we focus on those located in a one-to-many non target table, namely, those of table *Control*. This table contains 66 variables describing 10 610 tests performed on the 1 417 patients over 20 years of the study. Figure 7 depicts the STULONG dataset relational schema [6].

Since in this article we focus on binary variables, categorical variables are binarized. The purpose of the experiment is to find the most relevant secondary variables for predicting the target ones. The experimental results obtained on the STULONG dataset are presented in Table 1. We present for each target variable the five most relevant secondary variables. The results show that even by binarizing the categorical variables, we are able to identify those having a correlation with the target. Even by using merely a single secondary binary variable, the bivariate datagrid classifier offers high test accuracy and AUC. Table 2 shows some

[5] A cardiovascular disease due to the gradual accumulation of fatty materials.

[6] A description of all variables could be found on `http://euromise.vse.cz/challenge2004/data/index.html`

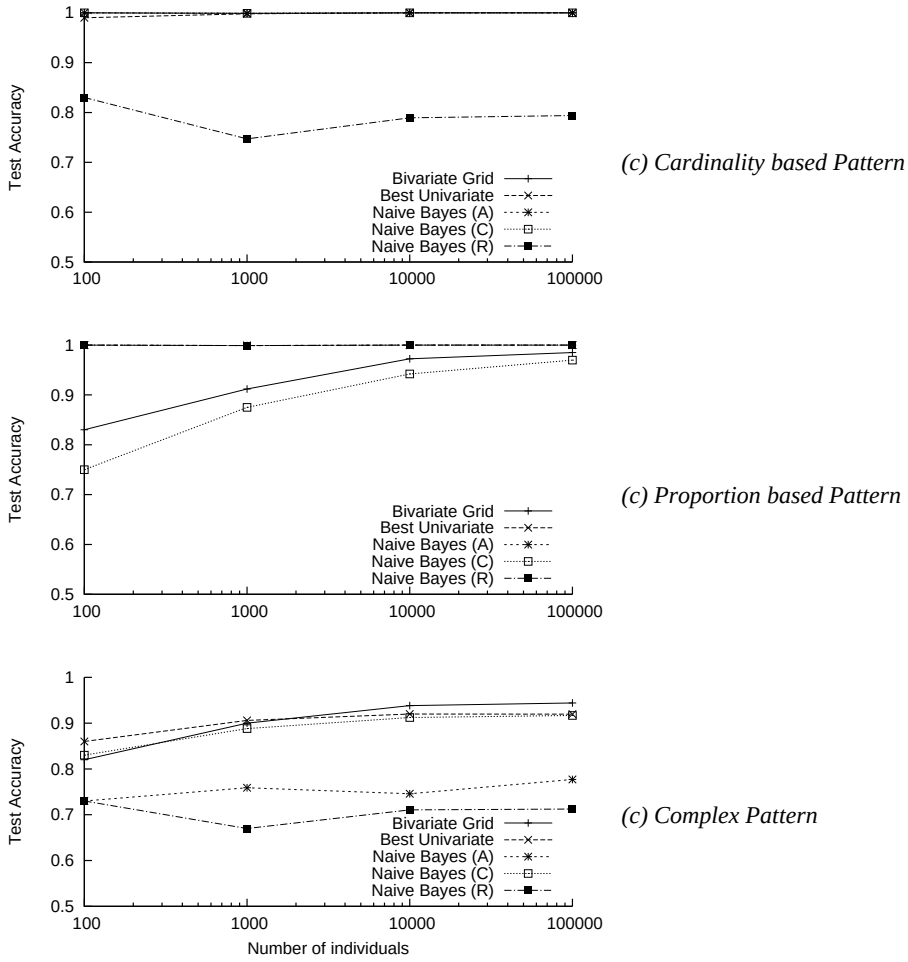


Fig. 6: Comparison between Bivariate Grid, Naive Bayes and Best Univariate Classifiers on artificial datasets

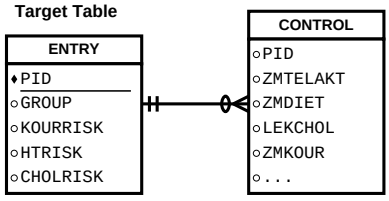


Fig. 7: Extract of the STULONG relational schema

Target Variable	Variable	test acc. (%)	test AUC (%)	Grid size
GROUP (0,69)*	HODNSK (B2)†	88,7 (±2, 1)	88,6 (±2, 9)	4
	HODNSK (B1)	82,4 (±2, 7)	84,3 (±3, 7)	6,6
	ZMKOUR (B0)	80,3 (±2, 7)	84,2 (±3, 3)	6
	HODNSK (B0)	82,9 (±2, 6)	83,4 (±3, 4)	6,2
	ZMKOUR (B1)	81,4 (±2, 9)	82,5 (±3, 0)	4
HTRISK (0,73)	LEKTLAK (B0)	81,2 (±1, 9)	79,1 (±3, 2)	6
	HYPERSD‡	83,2 (±1, 6)	78,0 (±3, 8)	4
	LEKTLAK (B2)	76,3 (±1, 4)	71,9 (±2, 7)	6
	LEKTLAK (B4)	77,7 (±1, 6)	68,7 (±3, 1)	4
	HYPCHL‡	74,0 (±1, 6)	65,7 (±2, 8)	6
KOURRISK (0,55)	ZMKOUR (B0)	81,3 (±4, 2)	85,0 (±4, 2)	6
	HODNSK (B2)	71,9 (±3, 6)	79,0 (±3, 8)	6
	ZMKOUR (B2)	75,1 (±5, 3)	78,5 (±5, 1)	6
	ZMKOUR (B3)	70,3 (±4, 9)	74,1 (±4, 5)	6
	HODNSK (B1)	65,0 (±2, 9)	72,4 (±3, 0)	6
CHOLRISK (0,72)	HYPCHL‡	83,5 (±3, 1)	79,2 (±4, 3)	4
	ZMDIET (B0)	72,8 (±0, 3)	63,1 (±4, 9)	4
	HODNSK (B2)	72,8 (±0, 3)	62,1 (±2, 7)	4
	JINAONE (B1)	72,8 (±0, 3)	60,4 (±3, 7)	4
	HYPTGL‡	72,8 (±0, 3)	59,9 (±3, 6)	4

* Percentage of the majority class

† The notation $(Bi)_{i\geq 0}$ denotes a binarized variable. for example, if HODNSK has 3 modalities then its binarization generates 3 features: HODNSK(B0), HODNSK(B1) and HODNSK(B2)

‡ These variables are originally binary

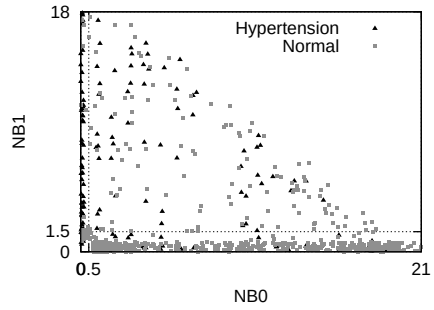
Table 1: Examples of relevant variables w.r.t some target variables in the STULONG datasets

correlations found between the studied target variables and some secondary variables in table *Control*. A secondary variable is considered to be correlated with the target one if it presents a correlation with at least one of the corresponding binarized variables.

Take for example the binary variable HYPERSD that describes for each exam whether the patient has a Systolic/Diastolic HYPERtension. This variable is relevant to predict the value of the target variable HTRISK (which indicates whether the patient has high blood pressure). The transformation of the variable HYPERSD generates two variables NB1 (the number of examinations with a positive HYPERSD) and NB0 (the number of examinations with a negative HYPERSD). Figure 8 shows the distribution of patients on the bivariate space NB1 $\times$ NB0. The optimal bivariate discretization of the two variables NB1 and NB0 is represented on the scatterplot. The corresponding data grid can separate the two classes (normal and high blood pressure). It achieves 83.2 % of test acc. and 78 % of AUC.

Sec. Var. \ Target Var.	Atherosclerosis (GROUP)	Hypertension risk (HTRISK)	Smoking risk (KOURRISK)	Cholesterol risk (CHOLRISK)
Systolic and Diastolic Hypertension (HYPERSD)		×	×	
Hypercholesterolemia (HYPCHL)	×	×		×
Hypertriglyceridemia (HYPTGL)	×			×
Changes of smoking (ZMKOUR)	×		×	×
Changes of diet (ZMDIET)				×
Change of physical activity (ZMTELAKT)			×	
Takes medicines for decreasing of blood pressure (LEKTLAK)		×		
Examination Diagnostic (HODNSK)	×	×		

Table 2: Some correlations found in the STULONG dataset between secondary variables and the target one



NB0×NB1	0	> 0
$]2; 18]$	$(97; 0.10; 0.90)^{\dagger}$	$(106; 0.65; 0.35)$
$[0; 1]$	$(270; 0.92; 0.08)$	$(384; 0.84; 0.16)$

† 97 individuals with 10% "Normal" and 90% "Hypertension"

Fig. 8: Scatterplot built from the binarization of the HYPERSD secondary variable (HTRISK as target variable). The optimal bivariate partition is represented on the scatterplot. The table to the right summarizes the frequency values of the variable target per grid cell

4 Conclusion

In this paper, we have proposed an approach for assessing the relevance of a binary secondary variable in the context of multi-relational supervised learning. The method consists of flattening the original relational representation by creating two numerical features which represent the cardinalities of the original binary values. A bivariate discretization model of both these features is generated, thereby inducing a bivariate partition. This partition qualifies the information provided jointly by the new features with respect to the target variable which is equivalent to assessing the relevance of the initial secondary variable since its information is not lost. For this, a criterion is proposed to evaluate each bivariate partition by means of a Bayesian approach.

We evaluated our approach on artificial and real datasets. Preliminary results on binary variables show that the evaluation criterion allows selecting highly informative variables. Future works are envisaged to provide efficient optimization procedures of this criterion in the case of continuous and categorical variables (eventually with large number of values). Classifiers using a univariate preprocessing like Naive Bayes or Decision Trees could therefore 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 Information Systems 22(1), pp. 2340 (2004)
2. Boullé, M.: MODL: A Bayes optimal discretization method for continuous attributes. Machine learning 65(1). pp. 131–165 (2006)
3. Boullé, M.: Recherche d'une représentation des données efficace pour la fouille des grandes bases de données. Ph.D. thesis, École Nationale Supérieure des Télécommunications (2007)
4. 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) pp. 39–61 (2009)
5. Džeroski, S.: Inductive logic programming and knowledge discovery in databases, pp. 117–152. American Association for Artificial Intelligence, Menlo Park, CA, USA (1996)
6. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Tech. Rep. HPL-2003-4, Hewlett Packard Laboratories (4 2003)
7. Frnkranz, J.: Dimensionality reduction in ILP : A call to arms. In L. De Raedt et S. Muggleton (Eds.), Proceedings of the IJCAI-97 Workshop on Frontiers of Inductive Logic Programming, pp. 81-86. Nagoya, Japan. (1997)
8. Guyon, I., A. Elisseeff : An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research 3*, pp. 1157–1182 (2003)
9. Hu, B., Liu, H., He, J., Du, X.: FARS: A Multi-relational Feature and Relation Selection Approach for Efficient Classification, pp. 73–86. Springer Verlag, Berlin, Germany (2008)
10. Knobbe, A.J., Blockeel, H., Siebes, A., Van Der Wallen, D.: Multi-Relational Data Mining. In: Proceedings of Benelearn '99 (1999)

11. Kramer, S., Flach,P. A., Lavrač, N.: Propositionalization approaches to relational data mining, Chapter 11, pp. 262-286. New York, NY, USA : Springer-Verlag (2001)
12. Langley, P., W. Iba, and K. Thompson: An analysis of Bayesian classifiers. In: 10th National Conference on Artificial Intelligence, pp. 223–228. San Jose, CA: AAAI Press. (1992)
13. Tomečková, M., Rauch, J., Berka, P.: STULONG - Data from a Longitudinal Study of Atherosclerosis Risk Factors. In P. Berka (Ed.), ECML/PKDD-2002 Discovery Challenge Workshop Notes, Helsinki, Finland. (2002)