Supervised Pre-processing of Numerical Variables for Multi-Relational Data Mining

Dhafer Lahbib, Marc Boullé, and Dominique Laurent

Abstract 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.

Key words: Supervised Learning; Multi-Relational Data Mining; One-to-many Relationship; Discretization; Variable Selection; Naive Bayes

D. Lahbib · M. Boullé
France Telecom R&D, 2 Avenue Pierre Marzin - 23300 Lannion - France e-mail: dhafer.lahbib,marc.boulle@orange.com

D.Lahbib · D. Laurent
ETIS-CNRS-Universite de Cergy Pontoise-ENSEA - 95000 Cergy Pontoise - France e-mail: dominique.laurent@u-cergy.fr

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 format. This structure may be naturally described using the relational formalism where each object (target table record) refers to one or more records in other tables (secondary tables) through a foreign key.

Example 1. In the context of the Customer Relationship Management (CRM) problem, Figure 1 shows an extract of a virtual CRM relational database schema. In this schema, the table *Customer* is the target table, whereas *Order* and *Service* are secondary tables related to *Customer* through the foreign key *CID*. In this context, the problem may be, for instance, to identify the customers likely to be interested in a certain product or service. This problem turns into a classification problem for which the target variable is the *Status* attribute, which denotes whether the customer has already ordered a particular product.

Learning from relational data has recently received increasing attention in the literature. The term Multi-Relational Data Mining (MRDM) was initially introduced by [Knobbe et al., 1999] to address novel knowledge discovery techniques from multiple relational tables. The common point between these techniques is that they need to transform the relational representation. In Inductive Logic Programming ILP [Džeroski, 1996], data is recoded as logic formulas. This causes scalability problems especially with large-scale data. Other methods called by Propositionalisation [Kramer et al., 2001] 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 the naturally compact initial representation is lost but there is a risk of introducing statistical bias because of potential dependencies between the newly added variables.

Although variable pre-processing is at the core of the majority of propositional (single table) Data Mining systems, it has received much less attention in MRDM. Pre-processing, including variable selection and discretization of numerical values, is of great importance particularly in Multi-Relational context. This step is justified not only to improve the accuracy but also to reduce the very large hypothesis spaces in MRDM. The difficulty when dealing with multiple table data arises from the presence of one-to-many associations. In the attribute-value mono table case, each individual has a single value per variable. While in multiple table setting, for a non target table variable, an individual may have a value list (eventually empty) of varying size.

Example 2. Referring back to Example 1, predicting whether the customer would be interested in a given product does not only depend on the information of that customer. Indeed, the other products ordered by this customer might be relavant, because, for instance, variables such as the product *Weight* or *Price* may present

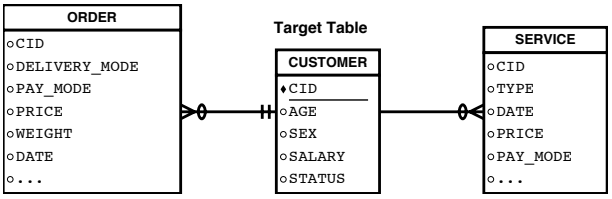


Fig. 1 Relational schema of a CRM database

correlations with the target variable. Assessing the relevance of these variables is not straightforward, since each customer may have made many orders. The same difficulty arises when trying to discretize accurately these numerical variables, especially when taking the class label into account.

To the best of our knowledge, only few studies in the literature have treated the numeric variable discretization in multi-relational data. Discretizing numerical attributes in multiple tables is different from handling attributes from a single table, due the presence of one-to-many associations. Under the multi-relational setting, the state of the art discretization approaches of a secondary numerical attribute differ along 2 axes: (i) whether they make use of the class label and (ii) whether they consider one-to-many relationships when computing cut points. The simplest methods that can be applied are the *equal-width* and *equal-frequency* interval binning. Both, are unsupervised and they compute boundaries regardless of any multi-relational structure. Whereas, the former divides the range of observed values into k equal sized bins, the latter discretizes the variable in such a way that each bin will have approximately the same number of values. To take into account the one-to-many association problem, [Knobbe and Ho, 2005] proposed an *Equal-weight* discretization method which involves an idea proposed by [Van Laer et al., 1997]: individuals with large number of related records in the non target table have a bigger influence on the choice of boundaries since they have more contributing numeric values. In order to compensate this impact, numeric values are weighted with the inverse of the size of the bags of records they belong to. Instead of producing ranges of equal size like in the equal frequency method, cut points are computed so that bins of equal weight can be obtained. All the above methods are class-blind since they do not use class labels. In order to take into account both the target variable information and the one-to-many association between records stored in the target and non-target tables, [Alfred, 2009] proposes a modification of the entropy-based multi-interval discretization method introduced by Fayyad and Irani [Fayyad and Irani, 1993]. Besides the class information entropy, another measure that uses individual information entropy is added to select multi-interval boundaries for the numerical secondary variable. The drawback of this approach is that it is relatively expensive and may lead to statistical skews since the entropy measures are computed by propagating the class labels to the non target tables. When performing such transformations, variables in the secondary table are not independent and identically distributed (i.i.d.).

In fact, individuals with a large number of related records in a secondary table will be overestimated thereby causing overfitting.

In this paper, we are interested in pre-processing a variable located in a secondary table having a one-to-many relation with the target one[1]. We propose to discretize the set of related values of a variable A and use an optimization criterion to find the best partitioning of the set such that the class Y is maximally differentiated. The idea is to use multi-variate data grids to estimate the conditional probability $P(Y \mid A)$. This univariate pre-processing extended to the relational context is of a great interest for filter feature selection [Guyon and Elisseeff, 2003] or as pre-processing step for classifiers such as Naive Bayes or Decision Tree.

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

2 Secondary Variables Pre-processing

In this section, we describe how a numerical variable belonging to a non-target table can be discretized in a class-dependent way.

2.1 Illustration of the Approach

Let us take the simplest case: a binary variable with two values v_1 and v_2. In this case, each individual is described by a bag of values among v_1 and v_2 [2]. Given an individual, all that we need to know about the secondary variable are the number of v_1 and the number of v_2 in the bag of records related to that individual (we denote them respectively n_1 and n_2). Thus, the whole information about the initial variable can be captured by considering jointly the pair (n_1, n_2). With such a representation, the conditional probability $P(Y \mid A)$ is then equivalent to $P(Y \mid n_1, n_2)$.

This approach can be generalized to a numerical secondary variable. In that case, the variable needs to be discretized into K intervals. The idea is to create in the target table K new variables n_k ($1 \leq k \leq K$). For each individual, n_k stands for the number of related records in the secondary table which have a value of A located in the k^{th} interval. As in the bivariate case, $P(Y \mid A)$ is approximated by evaluating $P(Y \mid (n_1, n_2, \cdots, n_K))$.

Multivariate data grid models have been shown to be good estimators for the probability of a class, given a set of input variables [Boullé, 2011]. The idea is to

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

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

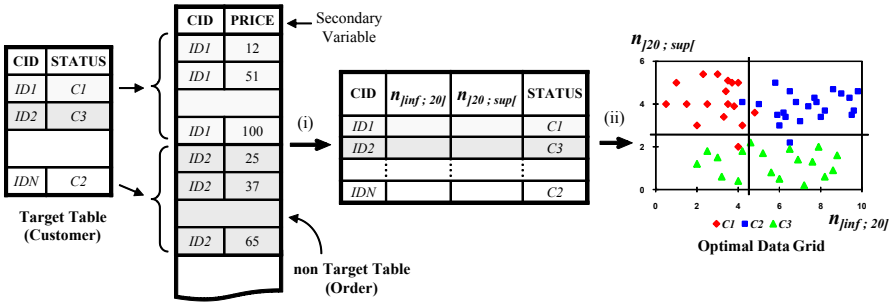


Fig. 2 Illustration of the Approach

jointly discretize in an optimal way the numeric variables n_k into intervals. This joint partitioning defines a distribution of the instances in a K-dimensional input data grid whose cells are defined by interval tuples. Therefore, our goal is to find the optimal multivariate discretization which maximizes the class distribution. In other words, we look for the optimal grid with homogeneous cells according to the class values.

Example 3. In the context of Example 1, consider, for instance, the secondary variable "PRICE" in the database of figure 1. Assume that we discretize this variable into two intervals : $]inf;20]$ and $]20;sup[$. Then PRICE is equivalent to the pair of variables $\left(n_{]inf;20]}, n_{]20;sup[}\right)$ where $n_{]inf;20]}$ (respectively $n_{]20;sup[}$) stands for the number of orders whose prices are less than 20 (respectively greater than 20). If we assume that the price is correlated with the target variable and that the discretization in two intervals is relevant, the target classes can be separated easily, using a grid similar to that of Figure 2.

The correlation between the cells of the data grid and the target values allows to quantify the joint classificatory information. The conditional probability distribution $P(Y \mid A)$ is evaluated locally in each cell. Consequently, classifiers like Naive Bayes or Decision Trees can easily be used. Moreover, it is important to note that the data grid provides an interpretable representation, since it shows the distribution of the individuals while jointly varying the count variables n_k. Each cell can be interpreted as a classification rule in the multi-relational context.

For example, the top-left cell of the data grid of Figure 2 is interpreted by: **if** "the number of orders with a price less than 20 is less than 5" **and** "the number of orders with a price greater than 20 is more than 2" **then** the class is $C1$.

Given that we use an equivalent representation, with the suitable discretization, we expect that the optimal related data grid will be able to detect the pattern contained in the secondary variable. Thus the problem is twofold: how to find the best discretization and how to optimize the related data grid. We address these two problems simultaneously by applying a model selection approach. To do so, we follow the MODL (Minimum Optimized Description Length) approach [Boullé, 2006]. The best model is chosen according to a Maximum A Posteriori (MAP) approach by maximizing the probability $p(\text{Model}|\text{Data})$ of the model given the data. By applying the Bayes rule, this is equivalent to maximizing $P(\text{Model})p(\text{Data}|\text{Model})$ since

the probability $P(\text{Data})$ is constant under varying the model. The considered models include the discretization of the secondary variable A and the joint partitioning of the generated count variables n_k. In the remainder of this section, we describe the criterion used to evaluate these models and we propose optimization algorithms.

2.2 Evaluation criterion

A model is completely defined by the discretization of the secondary variable (number and bounds of the intervals), the partitioning of the count variables n_k and the target distribution in each cell of the resulting data grid. To describe such a model, we use the following notation.

Notation 1

- *N : number of individuals (number of target table records)*
- *J : number of target values*
- *N_s : number of records in the non target table*
- *K : number of discretization intervals for the secondary variable A*
- *n_k : number of non target table records having a value of the secondary variable A in the k^{th} interval ($1 \leq k \leq K$)*
- *I_k : number of discretization intervals for the count variable n_k ($1 \leq k \leq K$)*
- *N_{i_k} : number of individuals in the interval i_k for variable n_k ($1 \leq k \leq K$)*
- *$N_{i_1 i_2 \ldots i_K}$: number of individuals in the cell $(i_1, i_2, \ldots, i_K)$*
- *$N_{i_1 i_2 \ldots i_K j}$: number of individuals in the cell $(i_1, i_2, \ldots, i_K)$ for the target value j*

Using the notation above, a model is completely defined by the parameters $\{K, \{n_k\}, \{I_k\}, \{N_{i_k}\}, \{N_{i_1 i_2 \ldots i_K j}\}\}$. In order to compute the criterion, we introduce in Definition 1 a prior distribution $p(\text{Model})$ on this model space. This prior makes explicitly the independence assumptions and exploits the hierarchy of the parameters. The number of discretization intervals of the secondary variable A is first chosen, then their bounds. After computing the count variables n_k, a K-dimensional data grid is built by choosing for each n_k the number of intervals, their bounds and finally the frequencies of the target values in each cell. At each stage of this hierarchy the choice is assumed to be uniform.

Definition 1 *The hierarchical prior of the parameters of discretization models is defined as follows:*

- *the numbers of intervals for the secondary variable discretization are independent from each other, and uniformly distributed between* 1 *and* N_s,
- *for a given number of intervals, every discretization of the secondary variable into intervals is equiprobable,*
- *for the discretization of the count variable n_k, the numbers of intervals are independent from each other, and uniformly distributed between* 1 *and* N,
- *for each count variable n_k and for a given number of intervals, every partition into intervals is equiprobable,*

- *for each cell of the data grid, all the parameters of the multinomial distribution of the target classes are equiprobable,*
- *the parameters of the multinomial distributions of the target classes in each cell are independent from each other.*

The first hypothesis of the above prior is that, for the secondary variable being discretized, the number of intervals is uniformly distributed between 1 and N_s. Thus we get

$$p(K) = \frac{1}{N_s} \tag{1}$$

The second hypothesis is that all discretizations of the secondary variable into K intervals are equiprobable for a given K. If N_s is the number of the secondary table records, there is $\binom{N_s + K - 1}{K - 1}$ ways to discretize N_s values into K intervals. Thus we obtain

$$p(\{n_k\} \mid K) = \frac{1}{\binom{N_s + K - 1}{K - 1}} \tag{2}$$

For each count variable n_k, the number of discretization intervals is uniformly distributed between 1 and N. Thus, we get

$$p(I_k \mid n_k, K) = \frac{1}{N} \tag{3}$$

For each count variable n_k, all the divisions of N instances into I_k intervals are equiprobable.

$$p\left(\left\{N_{i_k}\right\} \mid I_k, n_k, K\right) = \frac{1}{\binom{N + I_k - 1}{I_k - 1}} \tag{4}$$

Given K univariate discretizations of the count variables n_k, the frequency $N_{i_1 i_2 \ldots i_K j}$ of each cell $(i_1, i_2, \ldots, i_K)$ of the data grid can be derived from the input data sample. According to the fifth hypothesis of the prior distribution, in each cell $(i_1, i_2, \ldots, i_K)$, all the parameters of the multinomial distributions of the $N_{i_1 i_2 \ldots i_K}$. instances of the cell on the J target classes are equiprobable. Calculating the probability of a such set of multinomial parameters is a combinatorial problem, which turns into computing the number of ways of decomposing a natural number $N_{i_1 i_2 \ldots i_K}$ as a sum of J terms. Since each set of multinomial parameters is equiprobable, we obtain

$$p\left(\left\{N_{i_1 i_2 \ldots i_K j}\right\} \mid \left\{N_{i_k}\right\}, \{I_k\}, n_k, K\right) = \frac{1}{\binom{N_{i_1 i_2 \ldots i_K} + J - 1}{J - 1}} \tag{5}$$

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. This term is evaluated locally in each cell by considering the probability of observing the target values (classes) of the cell given the parameters of the multinomial distribution in this cell. The number of ways of observing $N_{i_1 i_2 \ldots i_K}$ instances distributed according to a multinomial distribution is given by the multinomial coefficient:

$$\frac{N_{i_1 i_2 \ldots i_K}!}{\prod_{j=1}^{J} N_{i_1 i_2 \ldots i_K j}!}.$$

The conditional likelihood per cell is thus

$$\frac{1}{\frac{N_{i_1 i_2 \ldots i_K}!}{\prod_{j=1}^{J} N_{i_1 i_2 \ldots i_K j}!}} \tag{6}$$

Taking the negative log of $P(\text{Model})p(\text{Data}|\text{Model})$, the generalized optimization criterion is given below.

$$\begin{aligned}
&\log N_s + \log \binom{N_s + K - 1}{K - 1} \\
&+ \sum_{k=1}^{K} \log N + \sum_{k=1}^{K} \log \binom{N + I_k - 1}{I_k - 1} \\
&+ \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_K=1}^{I_K} \log \binom{N_{i_1 i_2 \ldots i_K} + J - 1}{J - 1} \\
&+ \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_K=1}^{I_K} \left(\log N_{i_1 i_2 \ldots i_K}! - \sum_{j=1}^{J} \log N_{i_1 i_2 \ldots i_K j}! \right)
\end{aligned} \tag{7}$$

In formula 7, the first line stands for the choice of the discretization of the secondary variable: The first and second terms represent respectively the choices of the number of intervals, and the bounds of the intervals. The second and the third lines stand for the choice of the discretization of each variable n_k and the multinomial distribution parameters for the target values in each grid cell. The last term represents the conditional likelihood of the data given the model.

The criterion given in the above formula is related to the probability that the final data grid (obtained after the discretization of the secondary variable, as described before) explains the target variable given the secondary one. It can also be interpreted as the ability of a data grid to encode the target classes given the secondary variable, since negative log of probabilities is none other than a description length [Shannon, 1948].

Based on this cost, we can define a normalized compression gain $g(M)$ by considering the null model, denoted by $M_{\emptyset}$, where the secondary variable is discretized into one interval.

Algorithm 1: Optimization Algorithm

input $\mathscr{K}$: initial number of quantiles,
K_{max} : max number of evaluated ranges
output D^* : best secondary variable discretization,
G^* best Data Grid
require $K_{max} \ll \mathscr{K}$
Compute secondary variable quantiles bounds ($\mathscr{K}$-way equal frequency discretization) ;
Compute initial count variables $(v_k)_{1 \leq k \leq \mathscr{K}}$;
`/* Init solution (c*: best cost) */`
$c* \leftarrow \infty$, $D^* \leftarrow$ One interval, $G^* \leftarrow$ One cell;
for $K \leftarrow 2$ **to** K_{max} **do**
  $D \leftarrow$ discretize into K intervals;
  Estimate count variables $(n_k)_{1 \leq k \leq K}$ $n_k = \sum_{i=1+\left[\mathscr{K} \frac{k-1}{K}\right]}^{\left[\mathscr{K} \frac{k}{K}\right]} v_i$;
  Initialize G_K (data grid with n_k as input variables);
  `/* Optimize the data grid` G_K `*/`
  $G'_K \leftarrow OptimizeDataGrid(G_K)$;
  if $cost(G'_K) < c^*$ **then** `// if improved cost`
    `/* save improved solution */`
    $c^* \leftarrow cost(G'_K)$, $G^* \leftarrow G'_K$, $D^* \leftarrow D$;
  endif
endfor

$$g(M) = 1 - \frac{cost(M)}{cost(M_\emptyset)} \tag{8}$$

This relevance level can be used as a filter criterion for ranking secondary variables [Guyon and Elisseeff, 2003].

2.3 Optimization Algorithm

The choice of the secondary variable discretization is determined by the minimization of the criterion seen in Section 2.2, which is a combinatorial problem with 2^{N_s} possible discretizations for the secondary variable. Then, for each discretization into K intervals, there are $\left(2^N\right)^K$ possible data grids, which represent the number of the multivariate partitioning of the count variables $n_1, \ldots, n_K$. An exhaustive search through the whole space of models is unrealistic.

Algorithm 1 provides a simple procedure to optimize the discretization of the secondary variable. The method starts by making a fine $\mathscr{K}$-way equal frequency discretization of the secondary variable, which produces $\mathscr{K}$ initial count variables $v_1, \ldots, v_{\mathscr{K}}$. Then we iterate merging these initial ranges in order to simulate different equal frequency binnings. Each candidate discretization D_k is evaluated by optimizing the corresponding data grid G_K. This is done using the multivariate data

	# tables	# Numerical Sec. var.	# non target records	# Individuals	# target values
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

Table 1 Description of the used data sets

grid optimization heuristics detailed in [Boullé, 2011], which have practical scaling properties, with $O(N)$ space complexity and $O(N\sqrt{N}\log N)$ time complexity. At the end of Algorithm 1, we select the secondary discretization with the minimum evaluation cost (see criterion 7).

Although this simple algorithm clearly partially exploits the richness of the considered models, it is a good validation of the overall approach. As a priority for future work, we plan to extend this optimization procedure in order to better explore the search space and discover more complex discretization patterns.

3 Experiments

Our approach has been evaluated through its impact as a pre-processing step to a Naive Bayes (NB) classifier. In this multi-relational NB, for a given one-to-many numerical variable X_i, the optimal data grid gives an estimation of the corresponding univariate conditional density $P(X_i \mid Y)$, which is computed by considering the class frequencies in each cell. To show the contribution of our pre-processing approach over aggregation based methods, for each secondary attribute, the average value has been computed and a usual NB has been applied on the resulting flat table. Other aggregates were tested, namely Max, Min and the Number of records in secondary table. Results similar to those described below were obtained, and are omitted due to lack of space.

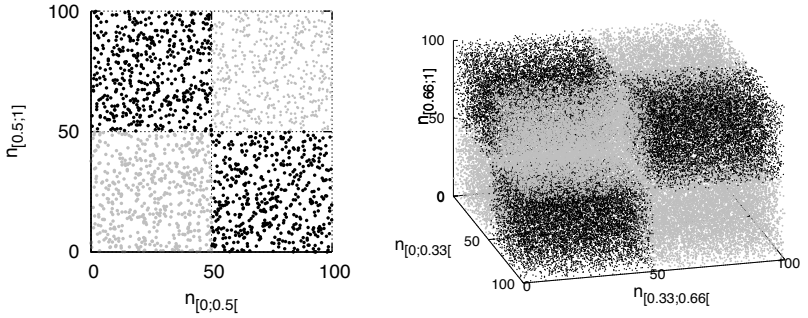
In our experiments, we have considered different classification tasks based on synthetic and real world data sets, whose characteristics are shown in Table 1.

[3] `http://sourceforge.net/projects/proper/files/datasets/0.1.0/`

[4] `http://cui.unige.ch/~woznica/rel_weka/`

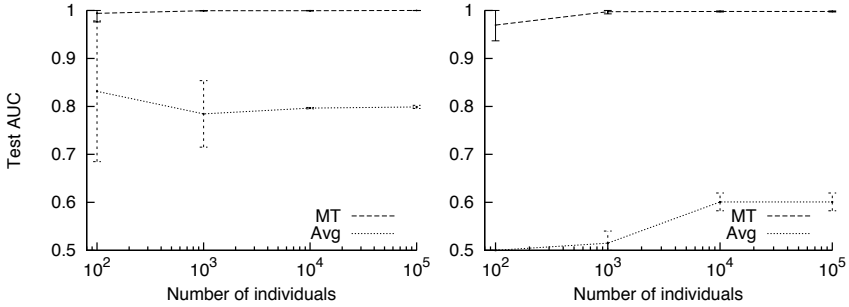
[5] `http://lamda.nju.edu.cn/data_MIMLimage.ashx`

[6] The study (STULONG) was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudík, MD, ScD, with collaboration of M. Tomečková, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. The data were transferred



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

Fig. 3 Scatter plots of synthetic data sets. Colors (black and gray) refer to the class labels



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

Fig. 4 Experimental results obtained on synthetic data sets

Regarding synthetic data sets, the ideal binning pattern is known in advance, and the target label is generated according to an Xor function between the count variables n_i. Figure 3 depicts the scatter plots of the 2D and 3D Xor datasets. For instance, in the 3D Xor pattern (Figure 3(b)), the secondary variable is supposed to be discretized into three intervals: $[0;0.33[$, $[0.33;0.66[$ and $[0.66;1[$. In this rather complex pattern, data points located, for example, at the corner near the origin (in gray) refer to individuals which have less than 50 values in the non target table, respectively, in the intervals $[0;0.63[$, $[0.33;0.66[$ and $[0.66;1[$.

To compare results, we recorded the Area Under the ROC Curve (AUC) using ten-fold cross-validation. The AUC criterion (see [Fawcett, 2003]) evaluates the ranking of the class conditional probabilities. In a two-class problem, the AUC is

to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head. Prof. RNDr. J. Zvárová, DrSc). The data resource is on the web pages `http://euromise.vse.cz/challenge2004`. At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107.

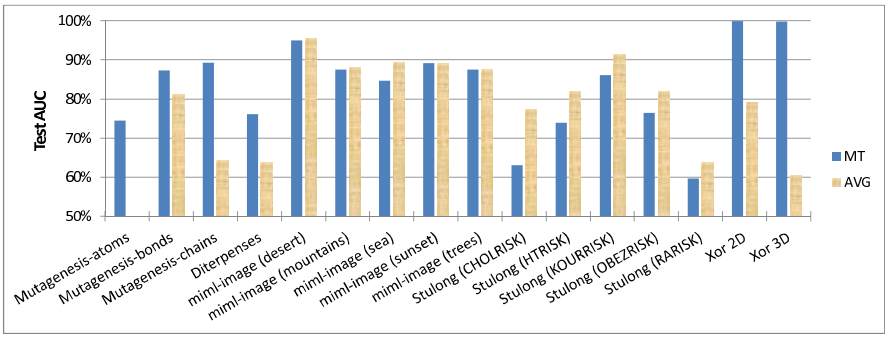


Fig. 5 Results of empirical data experiments obtained on artificial and real world data sets

equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In our experiments, we use the approach of [Provost and Domingos, 2001] to calculate the multi-class AUC, by computing each one-against-the-others two-classes AUC and weighting them by the class prior probabilities $P(Y_j)$.

In all experiment, only secondary numerical variables have been considered in the data sets and we have chosen $\mathscr{K} = 100$ and $K_{max} = 10$ as parameter for the optimization algorithm (cf. Algorithm 1). Obviously, it is not enough that only 10 equal frequency discretization are evaluated among $o\left(2^{N_s}\right)$ candidate discretization of the secondary variable. The objective of these experimentations is mainly to evaluate the potential of the approach, and investigate whether working on more sophisticated optimization algorithms is worth it.

Figure 5 shows the generalization performance (test AUC) obtained with a NB using our discretization approach (denoted MT) compared to the same classifier based on aggregated variables (denoted Avg). A two-tailed Student test at the 5% confidence level is performed in order to evaluate the significant wins or losses of our method versus the AVG method.

On synthetic data sets (Xor 2D and Xor 3D) our method widely outperforms the NB approach using the average value. Not surprisingly, this is explained by the fact that aggregation implies loss of information. On the other hand, our approach is able to recognize the pattern in the secondary variable and thus to discretize it correctly. This is confirmed by Figure 4, which summarizes the classification results obtained by varying the number of individuals in the artificial data sets. It can be seen that, with enough individuals, our approach reaches the theoretical performance. On the other hand, other experiments on a totally random pattern show that our method is robust, in the sense that it can detect the absence of predictive information in the secondary variable (which is materialized by a single interval discretization and an AUC near 50%).

On real world data sets, neither of the two methods dominates the other. Indeed, Figure 5 shows that: (i) our approach might perform better than the aggregation approach (Mutagenesis (atoms, bonds, chains) and Diterpenses), (ii) the two ap-

proaches might perform equivalently (Miml), and (iii) on Stulong data set, the aggregation approach might perform better than ours. This can be explained by the fact that our criterion needs a large number of individuals to recognize existing patterns (this has been shown in [Boullé, 2011], for a similar criterion in the case of a single table), whereas, as shown in Table 1, the used real world data sets are relatively small. Furthermore, we recall that Algorithm 1 is fairly simple and does not exploit the whole potential of the discretization criterion. Indeed, Algorithm 1 simulates an equal frequency discretization, meaning that many improvements can be brought to it. These results reported above are confirmed by the student's test in terms of significant wins, draws and loses of our method compared to the AVG method. The test showed 4 significant wins of our method on the Mutagenesis data set (atoms, bonds and chains) and Diterpenses, 4 draws on the Miml dataset (desert, mountains, sunset and trees) and six loses on Stulong (CHOLRISK, HTRISK, KOURRISK, OBEZRISK and RARISK) as well as Miml (sea).

We would like to emphasize that, although our approach does not always perform better than the average approach, this could be explained by insufficient exploration of the model space. Moreover, the approach is able to detect complex patterns (cf. Figure 3) that any aggregate approach can *not* discover. The obtained discretization yields rules that can be of interest to the user. On the other hand, it should be clear that aggregate methods can *not* produce such rules.

To see an example of how we can interpret the resulting discretization of a secondary variable, let us consider the Stulong data set (consisting of a target table Patient in a one-to-many relationship with a table Exam), along with the secondary numerical variable CHLSTMG that describes for each exam the cholesterol level (mg). It turns out that this variable is relevant to predict the value of the target variable CHOLRISK, which indicates whether the patient has high cholesterol risk according to the two target values: Normal and Risky. Applying Algorithm 1 in this case leads to a discretization of CHLSTMG into two intervals, namely $]inf, 228.5[$ and $[228.5, sup[$, and Figure 6 depicts the optimal data grid corresponding to this binning (histograms show the distribution of the target values in each cell). This table can be interpreted as a set of four classification rules, one for each cell.

For example the top-left cell is equivalent to the rule: **If** there are at least 3 examinations with a cholesterol level less than 228.5 mg **and** there is no examination with a cholesterol level higher than 228.5 mg **then** the class is Normal (meaning no cholesterol risk).

4 Conclusion

In this paper, we have presented a novel approach to discretize numerical variables in a multi-relational setting. Specifically, we propose to project numerical data in secondary tables on the target one by means of binning, and then for each individual, to count records in each interval. Additionally, we have seen how candidate discretizations can be evaluated in a class-dependent way. A criterion has been

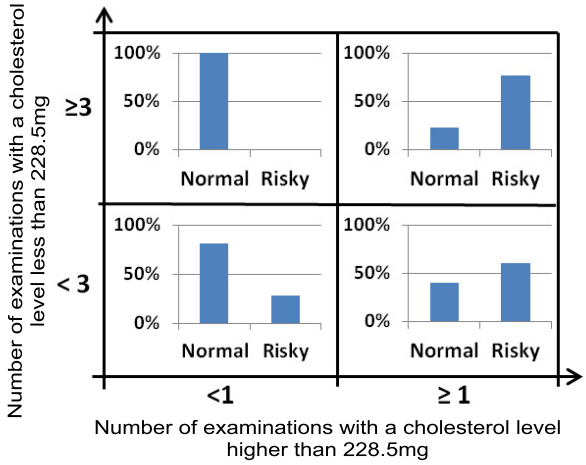


Fig. 6 Contingency table corresponding to the discretization of variable CHLSTMG

proposed to evaluate to what extent a given discretization of a secondary numerical variable preserves the correlation with the target variable. Finally, an optimization algorithm has been provided for computing the optimal discretization. We have shown that the criterion is robust and is able to evaluate a given discretization in a reliable way.

An algorithm has been given for computing an estimation of equal-frequency interval binning. This procedure, however, does not take full advantage from the potential of the criterion. We are currently investigating how to extend our algorithm in order to better explore the search space, so as to discover more accurate discretization patterns.

This study has shown, through experiments on artificial data sets, that the criterion and the discretization procedure may help in discovering relevant secondary variables and achieving high accuracy. However, in the case of real world data sets, we need to look for larger data sets, in order to better assess our approach and to compare it to other multi-relational data mining techniques.

References

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

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

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

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

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

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

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

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

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

[Kramer et al., 2001] Kramer, S., Flach, P. A., and Lavrač, N. (2001). Propositionalization approaches to relational data mining. In Džeroski, S. and Lavrač, N., editors, *Relational data mining*, chapter 11, pages 262–286. Springer-Verlag, New York, NY, USA.

[Provost and Domingos, 2001] Provost, F. and Domingos, P. (2001). Well-trained pets: Improving probability estimation trees. Technical Report CeDER #IS-00-04, New York University.

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

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