A Grouping Method for Categorical Attributes Having Very Large Number of Values

Marc Boullé

France Telecom R&D, 2, Avenue Pierre Marzin,
22300 Lannion, France
marc.boulle@francetelecom.com

Abstract. In supervised machine learning, the partitioning of the values (also called grouping) of a categorical attribute aims at constructing a new synthetic attribute which keeps the information of the initial attribute and reduces the number of its values. In case of very large number of values, the risk of overfitting the data increases sharply and building good groupings becomes difficult. In this paper, we propose two new grouping methods founded on a Bayesian approach, leading to Bayes optimal groupings. The first method exploits a standard schema for grouping models and the second one extends this schema by managing a "garbage" group dedicated to the least frequent values. Extensive comparative experiments demonstrate that the new grouping methods build high quality groupings in terms of predictive quality, robustness and small number of groups.

1 Introduction

Supervized learning consists in predicting the value of a class attribute from a set of explanatory attributes. Many induction algorithms rely on discrete attributes and need to discretize continuous attributes or to group the values of categorical attributes when they are too numerous. While the discretization problem has been studied extensively in the past, the grouping problem has not been explored so deeply in the literature. However, in real data mining studies, there are many cases where the grouping of values of categorical attributes is a mandatory preprocessing step. For example, most decision trees exploit a grouping method to handle categorical attributes, in order to increase the number of instances in each node of the tree. Neural nets are based on continuous attributes and often use a 1-to-N binary encoding to preprocess categorical attributes. When the categories are too numerous, this encoding scheme might be replaced by a grouping method. This problem arises in many other classification algorithms, such as bayesian networks or logistic regression. Moreover, the grouping is a general-purpose method that is intrinsically useful in the data preparation step of the data mining process [12].

When the categorical values are both few and highly informative, grouping the values might be harmful: the optimum is to do nothing, i.e. to produce one group per value. In case of very large number of categorical values, producing good groupings becomes harder since the risk of overfitting the data increases. In the limit situation

where the number of values is the same as the number of instances, overfitting is obviously so important that efficient grouping methods should produce one single group, leading to the elimination of the attribute. Many data mining commercial packages propose to eliminate attributes having too numerous values (for example, above a threshold of 100 values). While this is reliable, potentially informative attributes might be discarded. An efficient grouping method has to compromise between information and reliability, and determine the correct number of groups.

The grouping methods can be clustered according to the search strategy of the best partition and to the grouping criterion used to evaluate the partitions. The simplest algorithm tries to find the best bipartition with one category against all the others. A more interesting approach consists in searching a bipartition of all categories. The Sequential Forward Selection method derived from [6] and evaluated by [1] is a greedy algorithm that initializes a group with the best category (against the others), and iteratively adds new categories to this first group. When the class attribute has two values, [5] have proposed in CART an optimal method to group the categories into two groups for the Gini criterion. This algorithm first sorts the categories according to the probability of the first class value, and then searches for the best split in this sorted list. In the general case of more than two class values, there is no algorithm to find the optimal grouping with K groups, apart from exhaustive search. Decision tree algorithms often manage the grouping problem with a greedy heuristic based on a bottom-up classification of the categories. The algorithm starts with single category groups and then searches for the best merge between groups. The process is reiterated until no further merge improves the grouping criterion. The CHAID algorithm [7] uses this greedy approach with a criterion close to ChiMerge [8]. The best merges are searched by minimizing the chi-square criterion applied locally to two groups: they are merged if they are statistically similar. The ID3 algorithm [13] uses the information gain criterion to evaluate categorical attributes, without any grouping. This criterion tends to favor attributes with numerous categories and [14] proposed in C4.5 to exploit the gain ratio criterion, by dividing the information gain by the entropy of the categories. The chi-square criterion has also been applied globally on the whole set of categories, with a normalized version of the chi-square value [16] such as the Cramer's V or the Tschuprow's T, in order to compare two different-size partitions.

In this paper, we present a new grouping method called MODL, which results from a similar approach as that of the MODL discretization method [3]. This method is founded on a Bayesian approach to find the most probable grouping model given the data. We first define a general family of grouping models, and second propose a prior distribution on this model space. This leads to an evaluation criterion of groupings, whose minimization conducts to the optimal grouping. We use a greedy bottom-up algorithm to optimize this criterion. Additional preprocessing and post-optimization steps are proposed in order to improve the solutions while keeping a super-linear optimization time. The MODL method comes into a standard version where the grouping model consists of a partition of the categorical values, and into an extended version where a "garbage" group is settled to incorporate the least frequent values in a preprocessing step. Extensive experiments show that the MODL method produces high quality groupings in terms of compactness, robustness and accuracy.

The remainder of the paper is organized as follows. Section 2 describes the MODL method. Section 3 proceeds with an extensive experimental evaluation.

2 The MODL Grouping Method

In this section, we present the MODL approach which results in a Bayesian evaluation criterion of groupings and the greedy heuristic used to find a near Bayes optimal grouping.

2.1 Evaluation of a standard grouping model

The objective of the grouping process is to induce a set of groups from the set of values of a categorical explanatory attribute. The data sample consists of a set of instances described by pairs of values: the explanatory value and the class value. The explanatory values are categorical: they can be distinguished from each other, but they cannot *naturally* be sorted. We propose the following formal definition of a grouping model.

Definition 1: A *standard* grouping model is defined by the following properties:

1. the grouping model allows to define a partition of the categorical values into groups,
2. in each group, the distribution of the class values is defined by the frequencies of the class values in this group.

Such a grouping model is called a SGM model.

Notation:

n: number of instances
J: number of classes
I: number of categorical values
n_i: number of instances for value i
n_{ij}: number of instances for value i and class j
K: number of groups
$k(i)$: index of the group containing value i
n_k: number of instances for group k
n_{kj}: number of instances for group k and class j

The input data can be summarized knowing n, J, I and n_i. A SGM grouping model is completely defined by the parameters $\{ K, \{k(i)\}_{1 \le i \le I}, \{n_{kj}\}_{1 \le k \le K, 1 \le j \le J} \}$.

In the Bayesian approach, the best model is found by maximizing the probability $P(Model/Data)$ of the model given the data. Using the Bayes rule and since the probability $P(Data)$ is constant under varying the model, this is equivalent to maximizing $P(Model)P(Data/Model)$.

Once a prior distribution of the models is fixed, the Bayesian approach allows to find the optimal model of the data, provided that the calculation of the probabilities

$P(Model)$ and $P(Data/Model)$ is feasible. We present in Definiton 2 a prior which is essentially a uniform prior at each stage of the hierarchy of the model parameters. We also introduce a strong hypothesis of independence of the distribution of the class values. This hypothesis is often assumed (at least implicitly) by many grouping methods that try to merge similar groups and separate groups with significantly different distributions of class values. This is the case for example with the CHAID grouping method [7], which merges two adjacent groups if their distributions of class values are statistically similar (using the chi-square test of independence).

Definition 2: The following distribution prior on SGM models is called the three-stage prior:

1. the number of groups *K* is uniformly distributed between 1 and *I*,
2. for a given number of groups *K*, every division of the *I* categorical values into *K* groups is equiprobable,
3. for a given group, every distribution of class values in the group is equiprobable,
4. the distributions of the class values in each group are independent from each other.

Owing to the definition of the model space and its prior distribution, the Bayes formula is applicable to exactly calculate the prior probabilities of the model and the probability of the data given the model. Theorem 1, proven in [4], introduces a Bayes optimal evaluation criterion.

Theorem 1: A SGM model distributed according to the three-stage prior is Bayes optimal for a given set of categorical values if the value of the following criterion is minimal on the set of all SGM models:

$$\log(I)+\log(B(I,K))+\sum_{k=1}^{K}\log\left(C_{n_k+J-1}^{J-1}\right)+\sum_{k=1}^{K}\log\left(n_k!/n_{k,1}!n_{k,2}!..n_{k,J}!\right). \quad (1)$$

$B(I,K)$ is the number of divisions of the *I* values into *K* groups (with eventually empty groups). When $K=I$, $B(I,K)$ is the Bell number. In the general case, $B(I,K)$ can be written as a sum of Stirling numbers of the second kind:

$$B(I,K)=\sum_{k=1}^{K}S(I,k). \quad (2)$$

The first term of the criterion in Equation 1 stands for the choice of the number of groups, the second term for the choice of the division of the values into groups and the third term for the choice of the class distribution in each group. The last term encodes the probability of the data given the model.

2.2 Optimization of a standard grouping model

Once the optimality of an evaluation criterion is established, the problem is to design a search algorithm in order to find a grouping that minimizes the criterion. In this section, we present a standard greedy bottom-up heuristic. The method starts with initial single value groups and then searches for the best merge between groups. This merge is completed if it reduces the MODL evaluation criterion of the grouping and the process is reiterated until no further merge decreases the criterion.

With a straightforward implementation of the algorithm, the method runs in $O(n^3)$ time (more precisely $O(n+I^3)$). However, the method can be optimized in $O(n^2.log(n))$ time owing to an algorithm similar to that presented in [2]. The algorithm exploits the additivity of the evaluation criterion. Once a grouping is evaluated, the value of a new grouping resulting from the merge between two adjacent groups can be evaluated in a single step, without scanning all the other groups. Minimizing the value of the groupings after the merges is the same as maximizing the related variation of value Δvalue. These Δvalues can be kept in memory and sorted in a maintained sorted list (such as an AVL binary search tree for example). After a merge is completed, the Δvalues need to be updated only for the new group and its adjacent groups to prepare the next merge step.

Optimized greedy bottom-up merge algorithm:
- Initialization
  - Create an elementary group for each value: $O(n)$
  - Compute the value of this initial grouping: $O(n)$
  - Compute the Δvalues related to all the possible merges: $O(n^2)$
  - Sort the possible merges: $O(n^2.\log(n))$
- Optimization of the grouping
  Repeat the following steps: at most n steps
  - Search for the best possible merge: $O(1)$
  - Merge and continue if the best merge decreases the grouping value
    - Compute the Δvalues of the remaining group merges adjacent to the best merge: $O(n)$
    - Update the sorted list of merges: $O(n.\log(n))$

In the general case, the computational complexity is not compatible with large real databases, when the categorical values becomes too numerous. In order to keep a super-linear time complexity, we extend the greedy search algorithm with several preprocessing steps whose purpose is to reduce the initial number of categorical values. For example, "pure" values (related to one single class) can be merged with no degradation of the quality of the grouping. A more harmful heuristic consists in merging the least frequent values until the desired number of values is attained.

We also add some post-optimization heuristics to improve the final grouping solution. For example, every move of a categorical value from one group to another is evaluated and the best moves are performed as long as they improve the evaluation criterion. These additional pre-processing and post-optimization heuristics are detailed in [4].

2.3 The extended grouping model

When the number of categorical values increases, the grouping cost $B(I,K)$ in Equation 1 quickly rises and the potential group number falls down to 1. However, when the distribution of the categorical values is skewed, the most frequent values may be informative. A common practice in data preprocessing is to collect the least frequent values in a garbage group. In the extended grouping model presented in Definition 3, we generalize the standard grouping model by incorporating such a garbage group. After the preprocessing step, the remaining values are grouped using the standard model.

Definition 3: An *extended* grouping model is defined by the following properties:

1. the least frequent values are included into a special group called the *garbage* group,
2. the grouping model allows to define a partition of the remaining categorical values into groups,
3. in each group, the distribution of the class values is defined by the frequencies of the class values in this group.

Such a grouping model is called an EGM model.

Let F be the frequency threshold, such that the categorical values whose frequency is inferior to F are included in the garbage group. Let $I(F)$ be the number of remaining values (including the garbage group) once the preprocessing is performed. Although the extension increases the descriptive power of the model, we wish to trigger the extension only if necessary and to favor models close to the standard model, i.e. models with a small garbage frequency threshold. We express these prior preferences in Definition 4, using the universal prior for integers [15] for the distribution of F. Compared to the uniform prior, the universal prior for integers gives a higher probability to small integers with the smallest possible rate of decay. This provides a prior that favors models with small values of F.

The code length of the universal prior for integers is given by

$$L(n) = \log_2(c_0) + \log_2^*(n) = \log_2(c_0) + \sum_{j>1} \max\left(\log_2^{(j)}(n), 0\right), \tag{3}$$

where $\log_2^{(j)}(n)$ is the j[th] composition of $\log_2$ ($\log_2^{(1)}(n) = \log_2(n)$, $\log_2^{(2)}(n) = \log_2(\log_2(n))\ldots$) and $c_0 = \sum_{n>1} 2^{-\log_2^*(n)} = 2.865...$

Definition 4: The following distribution prior on EGM models is called the three-stage prior with garbage group:

1. using or not using a garbage group are two equiprobable choices,
2. the garbage frequency threshold F is distributed according the universal prior for integers,
3. the last parameters of the grouping model, with $I(F)$ categorical values, are distributed according the three stage prior.

Owing to this prior definition, we derive an evaluation criterion for the general grouping model in Theorem 2.

Theorem 2: An EGM model distributed according to the three-stage prior with garbage group is Bayes optimal for a given set of categorical values if the value of the following criterion is minimal on the set of all EGM models:

$$\log(2) + 1_{[2,\infty[}(F)L(F)\log(2) + \log(B(I(F),K)) + \sum_{k=1}^{K}\log\left(C_{n_k+J-1}^{J-1}\right) + \sum_{k=1}^{K}\log\left(n_k!/n_{k,1}!n_{k,2}!..n_{k,J}!\right). \quad (4)$$

The first term corresponds to the choice of using or not using a garbage group. The second term encodes the prior probability of the garbage frequency threshold, using the code length of the universal prior for integers. The last terms are those of the criterion presented in Theorem 1.

We now have to extend the search algorithm in order to find the most probable EGM model. A first step is to sort the explanatory values by increasing frequencies. This allows to quickly compute all possible frequency thresholds F and their corresponding remaining number of values $I(F)$. Once this step is completed, a basic algorithm consists in performing the standard search algorithm on SGM models for any frequency threshold F. In the worst case, this involves $O(\sqrt{n})$ runs of the standard search algorithm, since the number of distinct frequencies F (taken from the actual frequencies of the attribute values) cannot exceed $O(\sqrt{n})$ (their sum is bounded by n). The algorithm complexity of the extended search algorithm is thus $O(n\sqrt{n}\log(n))$.

In practice, the encoding cost of the garbage group is a minor part in the criterion presented in theorem 2. Introducing a garbage group becomes relevant only when a small increase of the frequency threshold brings a large decrease of the number of remaining categorical values. This property allows designing an efficient heuristic to find the garbage frequency threshold. This greedy heuristic first evaluates the simplest extended grouping (without garbage group) and then evaluates the extended groupings by increasing the garbage frequency threshold F as long as the criterion improves. Extensive experiments show that the practical complexity of the algorithms falls down to $O(n\log(n))$, with no significant decay in the quality of the groupings.

3 Experiments

In our experimental study, we compare the MODL grouping method with other supervised grouping algorithms. In this section, we introduce the evaluation protocol, the alternative evaluated grouping methods and the evaluation results.

3.1 The evaluation protocol

In order to evaluate the intrinsic performance of the grouping methods and eliminate the bias of the choice of a specific induction algorithm, we use a protocol similar as [2], where each grouping method is considered as an elementary inductive method.

We choose not to use the accuracy criterion because it focuses only on the majority class value and cannot differentiate correct predictions made with probability 1 from correct predictions made with probability slightly greater than 0.5. Furthermore, many applications, especially in the marketing field, rely on the scoring of the instances and need to evaluate the probability of each class value. To evaluate the predictive quality of the groupings, we use the Kullback-Leibler divergence [9] which compared the distribution of the class values estimated from the train dataset with the distribution of the class values observed on the test dataset. For a given categorical value, let p_j be the probability of the j^{th} class value estimated on the train dataset (on the basis of the group containing the categorical value), and q_j be the probability of the j^{th} class value observed on the test dataset (using directly the categorical value). The Kullback-Leibler divergence between the estimated distribution and the observed distribution is:

$$D(p \,||\, q) = \sum_{j=1}^{J} p_j \log \frac{p_j}{q_j} \quad . \tag{5}$$

The global evaluation of the predictive quality is computed as the mean of the Kullback-Leibler divergence on the test dataset. The q_j probabilities are estimated with the Laplace's estimator in order to deal with zero values.

The grouping problem is a bi-criteria problem that tries to compromise between the predictive quality and the number of groups. The optimal classifier is the Bayes classifier: in the case of an univariate classifier based on a single categorical attribute, the optimal grouping is to do nothing, *i.e.* to build one group per categorical value. In the context of data preparation, the objective is to keep most of the information contained in the attribute while decreasing the number of values. In the experiments, we collect both the predictive quality results using the Kullback-Leibler divergence and the number of groups.

In a first experiment, we compare the grouping methods considered as univariate classifiers. In a second experiment, we evaluate the results of the Naïve Bayes classifier using the grouping methods to preprocess the categorical attributes. In this experiment, the results are evaluated using the test accuracy and the robustness, computed as the ratio of the test accuracy by the train accuracy. We finally perform the same experiments using a Selective Naïve Bayes classifier.

We build a list of datasets having an increasing number of values per attribute on the basis of the Waveform dataset [5]. The Waveform dataset is composed of 5000 instances, 21 continuous attributes and a target attribute equidistributed on 3 classes. In order to build categorical attributes candidate for grouping, we discretize each continuous attribute in a preprocessing step with an equal-width unsupervised discretization,. We obtain a collection of 10 datasets using 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024 bin numbers for the equal-width algorithm. We build a second collection of "2D" datasets containing all the Cartesian products of the attributes.

Each of these 6 datasets (for bin numbers 2, 4, 8, 16, 32, 64) contains 210 categorical attributes. We finally produce a third collection of "3D" datasets on the basis of the Cartesian products of three attributes. Each of these 4 datasets (for bin numbers 2, 4, 8, 16) contains 1330 categorical attributes. On the whole, we get 20 datasets having a large variety of categorical attributes, with average number of values per attribute ranging from 2 to more than 1000.

3.2 The evaluated methods

The grouping methods studied in the comparison are:

- MODL: the extended MODL method described in this paper (using a garbage group),
- MODLS: the standard MODL method (without garbage group),
- CHAID [7],
- Tschuprow [16],
- Khiops [2],
- NoGrouping: one group per value.

All these methods are based on a greedy bottom-up algorithm that iteratively merges the categories into groups, and automatically determines the number of groups in the final partition of the categories. The MODL methods are based on a Bayesian approach and incorporate preprocessing and post-optimization algorithms. The CHAID, Tschuprow and Khiops methods exploit the chi-square criterion in different manner. The CHAID method is the grouping method used in the CHAID decision tree classifier. It applies the chi-square criterion locally to two rows of the contingency table, and iteratively merges the values as long as they are statistically similar. The significance level is set to 0.95 in the experiments. The Tschuprow method is based on a global evaluation of the contingency table, and uses the Tschuprow's T normalization of the chi-square value to evaluate the partitions. The Khiops method also applies the chi-square criterion on the whole contingency table, but it evaluates the partition using the confidence level related to the chi-square criterion instead of the Tschuprow criterion. It unconditionally groups the least frequent values in a preprocessing step in order to improve the reliability of the confidence level associated with the chi-square criterion, by constraining every cell of the contingency table to have an expected value of at least 5. Furthermore, the Khiops method provides a guaranteed resistance to noise: any categorical attribute independent from the class attribute is grouped in a single terminal group with a user defined probability. This probability is set to 0.95 in the experiments.

3.3 The univariate experiment

The goal of the univariate experiment is to evaluate the intrinsic performance of the grouping methods, without the bias of the choice of a specific induction algorithm. The grouping are performed on each attribute of the 20 synthetic datasets derived from the Waveform dataset, using a stratified tenfold cross-validation.

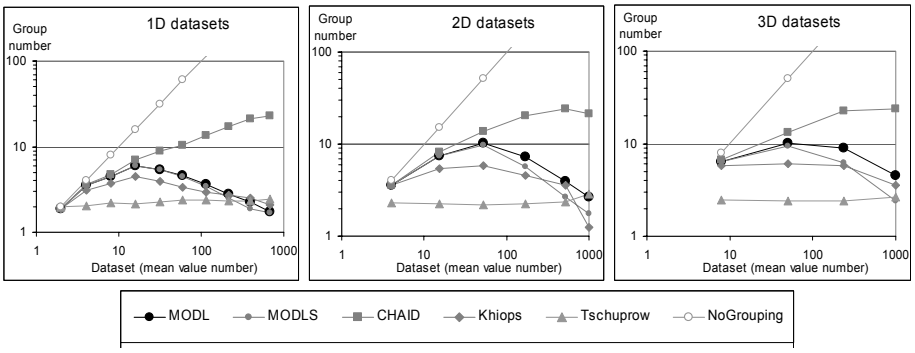

Fig. 1. Mean of the group number per attribute on the 20 datasets

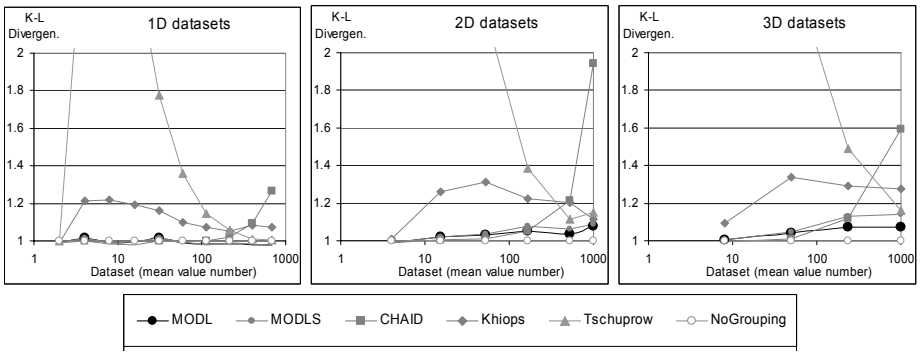

Fig. 2. Mean of the normalized Kullback-Leibler divergence per attribute on the 20 datasets

During the experiments, we collect the group number and the Kulback-Leibler divergence between the class distribution estimated on train datasets and the class distribution observed on test datasets. For each grouping method, this represents 210 measures for every 1D dataset, 2100 measures for every 2D dataset and 13300 for every 3D dataset. These results are summarized across the attributes of each dataset owing to means, in order to provide a gross estimation of the relative performances of the methods. We report the mean of the group number and of the Kullback-Leibler divergence for each dataset in Figures 1 and 2. The dataset result points are ordered by increasing bin number (from 2 bins to 1024 bins for the 1D datasets, from 2 bins to 64 bins the 2D datasets and from 2 bins to 16 bins for the 3D datasets). The result points are scaled on the x-coordinate according to the mean value number per attribute in each dataset, in order to visualize the relation between the number of values and the evaluated criterion. For the Kullback-Leibler divergence, we normalize each result by that of the NoGrouping method.

As expected, the NoGrouping method obtains the best results in term of predictive quality, at the expense of the worst number of groups. The Tschuprow method is heavily biased in favor of number of groups equal to the number of class values: it always produces between 2 and 3 groups, and obtains a very poor estimation of the

class distribution (evaluated by the Kullback-Leibler divergence) as shown in Figure 2. The Khiops method suffers from its minimum frequency constraint. It produces few groups and gets a reliable estimation of the class distribution across all the datasets, whatever their mean value number per attribute. However, it fails to obtain the best groupings on most of the datasets. The CHAID and MODL methods almost reach the predictive quality of the NoGrouping method with much smaller number of groups when the mean value number is less than 100. The CHAID method produces an increasing number of groups when the number of values rises. When the number of values if very large (between 100 and 1000), it overfits the data with too many groups, and its estimation of the class distribution worsen sharply as shown in Figure 2. The MODL methods always get the best estimation of the class distribution, very close to that of the NoGrouping method. They produce an increasing number of groups when the number of values is below a few tenths and then slowly decrease the number of groups. There is only a slight difference between the standard and the extended versions of the MODL method. When the number of values becomes very large, the extended version produces some extra groups owing to its garbage group and better approximates the class distribution.

To summarize, the MODL methods manage to get the lowest number of group without discarding the predictive quality.

3.4 The Naïve Bayes experiment

The aim of the naïve Bayes experiment is to evaluate the impact of grouping methods on the Naïve Bayes classifier. The Naïve Bayes classifier [10] assigns the most probable class value given the explanatory attributes values, assuming independence between the attributes for each class value. The probabilities for categorical attributes are estimated using the Laplace's estimator directly on the categorical values. The results are presented in Figure 3 for the test accuracy and in Figure 4 for the robustness (evaluated as the ratio of the test accuracy by the train accuracy).

Most methods do not perform better than the NoGrouping method. This probably explains why the Naïve Bayes classifiers do not make use of groupings in the literature. The Tschuprow method is hampered by its poor estimation of the class distribution and obtains test accuracy results that are always dominated by the NoGrouping method. The Khiops method obtains good accuracy and robustness results when the number of values is below 100. For higher numbers of values, it suffers from its minimum frequency constraint and its accuracy results dramatically fall down to the accuracy of the majority classifier (33% in the Waveform dataset). The CHAID method obtains results very close to the NoGrouping method, both on the accuracy and robustness criteria. The MODL methods clearly dominate all the other methods when the two criteria are considered. On the accuracy criterion, they obtain almost the same results than the CHAID and NoGrouping methods. On the robustness criterion, they strongly dominate these two methods. Once again, there is only a minor advantage for the extended version of the MODL method compared to its standard version.

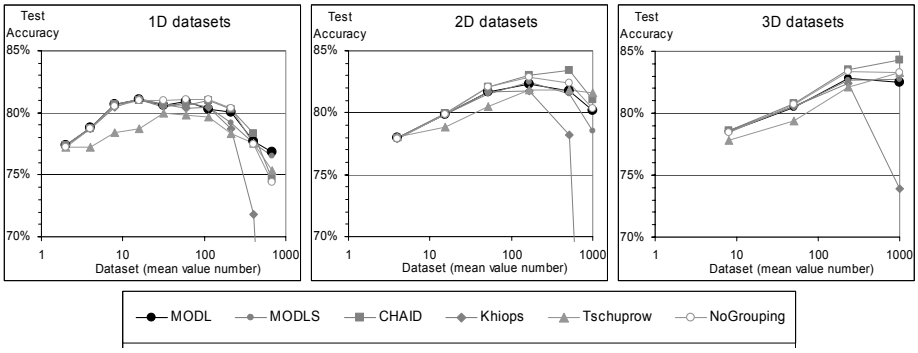

Fig. 3. Mean of the Naïve Bayes test accuracy on the 20 datasets

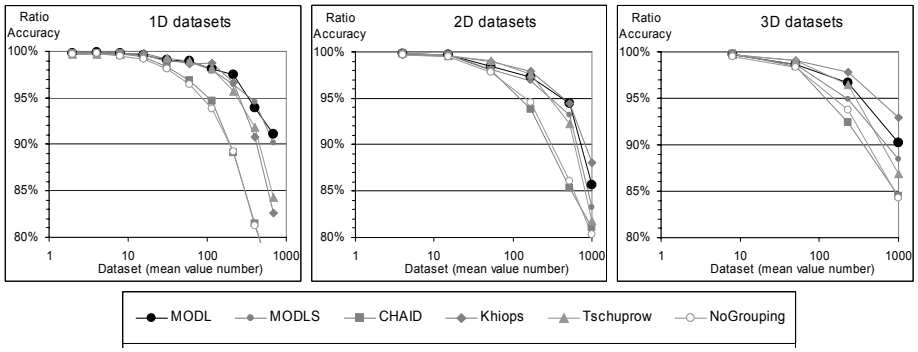

Fig. 4. Mean of the Naïve Bayes robustness on the 20 datasets

It is interesting to notice that the naïve Bayes classifier is very robust and manages to produce accurate predictions even in case of attributes having very large numbers of values. Another attractive aspect learnt from this experiment is the overall gain in test accuracy when the pairs (2D datasets) and triples (3D datasets) of attributes are considered. Using Cartesian products allows to investigate simple interactions between attributes and to go beyond the limiting independence assumption of the Naïve Bayes classifier. Although this degrades the robustness (because of a decrease in the frequency of the categorical values), this enhances the test accuracy.

3.5 The Selective Naïve Bayes experiment

The selective naïve Bayes classifier [11] incorporates feature selection in the naïve Bayes algorithm, using a stepwise forward selection. It iteratively selects the attributes as long as there is no decay in the accuracy. We use a variant of the evaluation and stopping criterion: the area under the lift curve instead of the accuracy. The lift curve summarizes the cumulative percent of targets recovered in the top quantiles of the sample [17]. The lift curve based criterion allows a more subtle

evaluation of the conditional class density than the accuracy criterion, which focuses only on the majority class.
Compared to the naïve Bayes (NB) classifier, the selective naïve Bayes (SNB) classifier is able to remove independent or redundant attributes owing to its selection process. However, it is more likely to overfit the data and requires a better evaluation of the predictive influence of each attribute. The purpose of the SNB experiment is to evaluate the impact of grouping on a classifier using an attribute selection process. The results are presented in Figure 5 for the test accuracy. The robustness results, not presented here, are very similar to those of the naïve Bayes experiment.

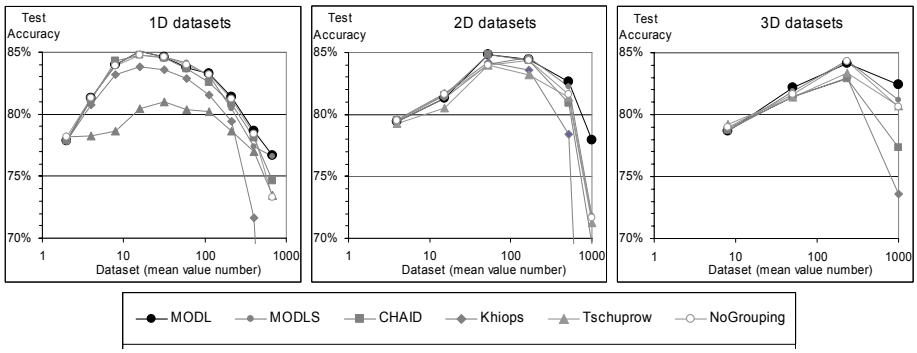

Fig. 5. Mean of the Selective Naïve Bayes test accuracy on the 20 datasets

The Tschuprow and Khiops grouping methods suffer from their respective limitations (strong bias and minimum frequency constraint): they are constantly dominated by the other methods. The MODL, CHAID and NoGrouping achieve comparable accuracy results when the mean value number is below 100. Above this threshold, the accuracy results decrease as the mean value number still increases. The CHAID method exhibits the worst rate of decrease, followed by the NoGrouping and finally the MODL methods. The extended MODL method always gets the best results. However, the benefit of the extended MODL method over the standard MODL method is still insignificant, except in the extreme case where the mean value number is close to 1000. For example, in the dataset (2D, 64 bins), the extended MODL method obtains a 77% test accuracy, about 6% above that of the standard MODL and NoGrouping methods and 8% above the CHAID method.

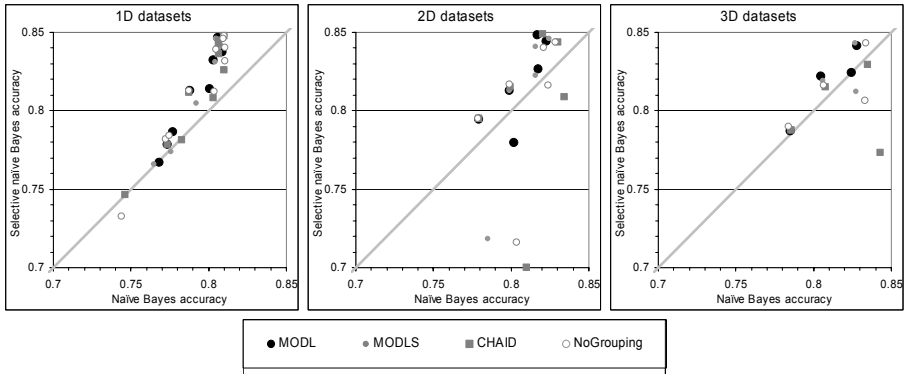

Fig. 6. Naïve Bayes versus Selective Naïve Bayes test accuracy on the 20 datasets

Apart from the grouping analysis, it is interesting to compare the results of the naïve Bayes and selective Bayes classifiers. Figure 6 reports the NB test accuracy per dataset on the x-coordinate and the SNB test accuracy per dataset on the y-coordinate for the most accurate grouping methods. Whereas the NB classifier obtains better accuracy results when pairs or triples of attributes are considered, this not the case for the SNB classifier. The SNB classifier applies its selection process to a larger set of attributes. This increases the risk of overfitting the data, so that the SNB classifier is not able to benefit from the additional information brought by the Cartesian products of attributes. On the opposite, for a given set of attributes, the SNB classifier almost always achieves better accuracy results than the NB classifier, especially with the extended MODL algorithm. Using this grouping method, the SNB classifier improves the NB classifier accuracy results on all the 20 datasets except one (2D, 64 bins). On a whole, the extended MODL method achieves the best results with the smallest variance across the datasets.

4 Conclusion

The MODL grouping methods exploits a precise definition of a family of grouping models with a general prior. This provides a new evaluation criterion which is minimal for the Bayes optimal grouping, *i.e.* the most probable grouping given the data sample. Compared to the standard version of MODL method, the extended version incorporates a garbage group dedicated to the least frequent values.

Extensive evaluations have been performed on a collection of datasets composed of varying numbers of attributes and mean numbers of values per attribute. The most difficult dataset consists of about 5000 instances and 1000 categorical attributes, each one having 1000 values. The experiments demonstrate that the MODL methods are very efficient: they build groupings that are both robust and accurate. Compared to the CHAID method, they reduce the number of groups by up to one order of magnitude and improve the estimation of the conditional class density. They allow classifiers to take benefit of informative attributes even when their numbers of values are very large, especially with the extended version of the MODL method.

References

1. Berckman, N.C.: Value grouping for binary decision trees. Technical Report, Computer Science Department – University of Massachusetts (1995)
2. Boullé, M.: A robust method for partitioning the values of categorical attributes. Revue des Nouvelles Technologies de l'Information, Extraction et gestion des connaissances (EGC'2004), RNTI-E-2, volume II, (2004a) 173-182
3. Boullé, M.: A Bayesian Approach for Supervised Discretization, Data Mining V, Eds Zanasi, Ebecken, Brebbia, WIT Press, (2004b) 199-208
4. Boullé, M.: MODL: une méthode quasi-optimale de groupage des valeurs d'un attribut symbolique. Note Technique NT/FT/R&D/8611. France Telecom R&D (2004c)
5. Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J.: Classification and Regression Trees. California: Wadsworth International (1984)
6. Cestnik, B., Kononenko, I. & Bratko, I.: ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In Bratko & Lavrac (Eds.), Progress in Machine Learning. Wilmslow, UK: Sigma Press, (1987)
7. Kass, G.V.: An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29(2) (1980) 119-127
8. Kerber, R.: Chimerge discretization of numeric attributes. Proceedings of the 10th International Conference on Artificial Intelligence (1991) 123-128
9. Kullback, S.: Information Theory and Statistics. New York: Wiley, (1959); republished by Dover, (1968)
10. Langley, P., Iba, W., & Thompson, K.: An analysis of bayesian classifiers. In Proceedings of the 10th national conference on Artificial Intelligence, AAAI Press, (1992) 223-228
11. Langley, P., & Sage, S.: Induction of Selective Bayesian Classifiers. In Proc. of the 10th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann (1994) 399-406
12. Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann (1999)
13. Quinlan, J.R.: Induction of decision trees. Machine Learning, 1, (1986) 81-106
14. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
15. Rissanen, J.: A universal prior for integers and estimation by minimum description length. Ann. Statis. 11 (1983) 416-431
16. Ritschard, G., Zighed, D.A. & Nicoloyannis, N.: Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé. Math. & Sci. Hum., n°154-155, (2001) 81-98
17. Witten, I.H. & Franck, E.: Data Mining. Morgan Kaufmann (2000)