

A Bayes Optimal Approach for Partitioning the Values of Categorical Attributes

Marc Boullé

*France Telecom R&D
2 Avenue Pierre Marzin
22300 Lannion, France*

MARC.BOULLE@FRANCETELECOM.COM

Editor: Greg Ridgeway

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 this paper, we propose a new grouping method MODL¹ founded on a Bayesian approach. The method relies on a model space of grouping models and on a prior distribution defined on this model space. This results in an evaluation criterion of grouping, which is minimal for the most probable grouping given the data, *i.e.* the Bayes optimal grouping. We propose new super-linear optimization heuristics that yields near-optimal groupings. Extensive comparative experiments demonstrate that the MODL grouping method builds high quality groupings in terms of predictive quality, robustness and small number of groups.

Keywords: data preparation, grouping, Bayesianism, model selection, classification, naïve Bayes

1 Introduction

Supervised learning consists of 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. The grouping problem consists in partitioning the set of values of a categorical attribute into a finite number of groups. 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 (Zighed and Rakotomalala, 2000). 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 (Pyle, 1999).

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 that of Cestnik et al. (1987) and evaluated by Berckman (1995) 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, Breiman et al. (1984) have proposed in CART an optimal method to group the categories into two groups for the Gini criterion. This

¹ This work is covered under French patent number 04 00179.

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. This algorithm has a time complexity of $O(I \log(I))$, where I is the number of categories. Based on the ideas presented in (Lechevallier, 1990; Fulton et al., 1995), this result can possibly be extended to find the optimal partition of the categories into K groups in the case of two class values, with the use of a dynamic programming algorithm of time complexity I^2 . 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. However, Chou (1991) has proposed an approach based on K-means that allows finding a locally optimal partition of the categories into K groups. 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 can improve the grouping criterion. The CHAID algorithm (Kass, 1980) uses this greedy approach with a criterion close to ChiMerge (Kerber, 1991). The best merges are searched by minimizing the chi-square criterion applied locally to two categories: they are merged if they are statistically similar. The ID3 algorithm (Quinlan, 1986) uses the information gain criterion to evaluate categorical attributes, without any grouping. This criterion tends to favor attributes with numerous categories and Quinlan (1993) 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 (Ritschard et al., 2001) 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 (Boullé, 2004c). 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 defines the optimal grouping. We use a greedy bottom-up algorithm to optimize this criterion. The method starts the grouping from the elementary single value groups. It evaluates all merges between groups, selects the best one according to the MODL criterion and iterates this process. As the grouping problem has been turned into a minimization problem, the method automatically stops merging groups as soon as the evaluation of the resulting grouping does not decrease anymore. Additional preprocessing and post-optimization steps are proposed in order to improve the solutions while keeping a super-linear optimization time. 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 Bayes optimal evaluation criterion of groupings and the greedy heuristic used to find a near-optimal grouping.

2.1 Presentation

In order to introduce the issues of grouping, we present in Figure 1 an example based on the Mushroom UCI data set (Blake and Merz, 1998). The class attribute has two values: EDIBLE and POISONOUS. The 10 categorical values of the explanatory attribute CapColor are sorted by decreasing frequency; the proportions of the class values are reported for each explanatory value. Grouping the categorical values does not make sense in the unsupervised context. However, taking the class attribute into account introduces a metric between the categorical values. For example, looking at the proportions of their class values, the YELLOW cap looks closer from the RED cap than from the WHITE cap.

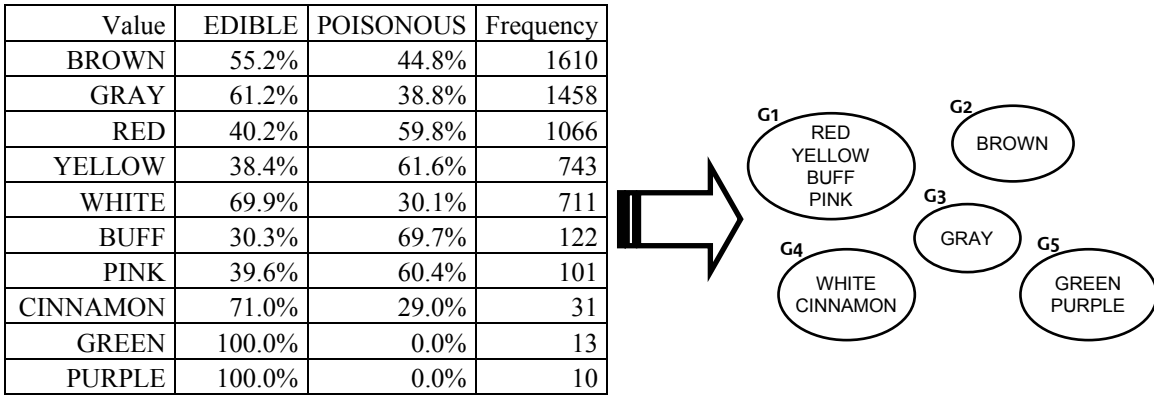


Figure 1. Example of a grouping of the categorical values of the attribute CapColor of data set Mushroom

In data preparation for supervised learning, the problem of grouping is to produce the smallest possible number of groups with the slightest decay of information concerning the class values. In Figure 1, the values of CapColor are partitioned into 5 groups. The BROWN and GRAY caps are kept into 2 separate groups, since their relatively small difference of proportions of the class values is significant (both categorical values have important frequencies). On the opposite, the BUFF cap is merged with the RED, YELLOW and PINK caps: the frequency of the BUFF cap is not sufficient to make a significant difference with the other values of the group.

The issue of a good grouping method is to find a good trade-off between information (as many groups as possible with discriminating proportions of class values) and reliability (the class information learnt on the train data should be a good estimation of the class information observed on the test data). Producing a good grouping is harder with large numbers of values 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. In real applications, there are some domains that require grouping of the categorical attributes. In marketing applications for example, attributes such as Country, State, ZipCode, FirstName, ProductID usually hold many different values. Preprocessing these attributes is critical to produce efficient classifiers.

2.2 Definition of a 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 in Definition 1 the following formal definition of a grouping model. Such a model is a pattern that describes both the partition of the categorical values into groups and the proportions of the class values in each group.

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

1. the grouping model allows to describe 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.

Notations:

- 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 purpose of a grouping model is to describe the distribution of the class attribute conditionally to the explanatory attribute. All the information concerning the explanatory attribute, such as the size of the data set, the number of explanatory values and their distribution might be used by a grouping model. The input data can be summarized knowing n , J , I and n_i . The grouping model has to describe the partition of the explanatory values into groups and the distribution of the class values in each group. A SGM grouping model is completely defined by the parameters $\{ K, \{k(i)\}_{1 \leq i \leq I}, \{n_{kj}\}_{1 \leq k \leq K, 1 \leq j \leq J} \}$.

For example, in Figure 1, the input data consists of $n=5865$ instances (size of the train data set used in the sample), $I=10$ categorical values and $J=2$ class values. The n_i parameters represents the counts of the categorical values (for example, $n_1=1610$ for the BROWN CapColor).

The grouping model pictured in Figure 1 is defined by $K=5$ groups, the description of the partition of the 10 values into 5 groups (for example, $k(1)=2$ since the BROWN CapColor belongs to the G2 group) and the description of the distribution of the class values in each group. This last set of parameters ($\{n_{kj}\}_{1 \leq k \leq 5, 1 \leq j \leq 2}$) corresponds to the counts in the contingency table of the grouped attribute and class attribute.

2.3 Evaluation of a Grouping Model

In the Bayesian approach, the best model is found by maximizing the probability $P(\text{Model}/\text{Data})$ of the model given the data. Using Bayes rule and since the probability $P(\text{Data})$ is constant under varying the model, this is equivalent to maximize $P(\text{Model})P(\text{Data}/\text{Model})$. For a detailed presentation on Bayesian theory and its applications to model comparison and hypothesis testing, see for example (Bernardo and Smith, 1994; Kass and Raftery, 1995).

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 $P(\text{Model})$ and $P(\text{Data}/\text{Model})$ is feasible. We define below 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 (Kass, 1980), 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, 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 Appendix A, 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:

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

C is the combinatorial operator. $B(I, K)$ 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 $S(I, k)$:

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

$S(I, k)$ stands for the number of ways of partitioning a set of I elements into k nonempty sets.

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.

There is a subtlety in the three-stage prior, where choosing a grouping with K groups incorporates the case of potentially empty groups. The partitions are thus constrained to have at most K groups instead of exactly K groups. The intuition behind the choice of this prior is that if K groups are chosen and if the categorical values are dropped in the groups independently from each other, empty groups are likely to appear. We present below two theorems (proven in Appendix A) that bring a more theoretical justification of the choice of the prior. These theorems are no longer true when the prior is to have partition containing exactly K groups.

Definition 3: A categorical value is *pure* if it is associated with a single class.

Theorem 2: In a Bayes optimal SGM model distributed according to the three-stage prior, two pure categorical values having the same class are necessary in the same group.

This brings an intuitive validation of the MODL approach. Furthermore, grouping algorithms can exploit this property in a preprocessing step and considerably reduce their overall computational complexity.

Theorem 3: In a SGM model distributed according to the three-stage prior and in the case of two classes, the Bayes optimal grouping model consists of a single group when each instance has a different categorical value.

This provides another validation of the MODL approach. Building several groups in this case would reflect an over-fitting behavior.

Conjecture 1: In a Bayes optimal SGM model distributed according to the three-stage prior and in the case of two classes, any categorical value whose class proportion is between the class proportions of two categorical values belonging to the same group necessary belongs to this group.

This conjecture has been proven for other grouping criterion such as Gini (Breiman, 1984) or Kolmogorov-Smirnov (Asseraf, 2000) and experimentally validated in extensive experiments for the MODL criterion. It will be considered as true in the following. The grouping algorithms, such as greedy bottom-up merge algorithms, can take benefit from this conjecture. Once the categorical values have been sorted by decreasing proportion of the class, the number of potentially interesting value merges is reduced to $(I-1)$ instead of $I(I-1)/2$.

2.4 Optimization of a 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 model that minimizes the criterion. In this section, we present a greedy bottom-up merge heuristic enhanced with several preprocessing and post-optimization algorithms whose purpose is to achieve a good trade-off between the time complexity of the search algorithm and the quality of the groupings.

2.4.1 Greedy Bottom-Up Merge Heuristic

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 performed if the MODL evaluation criterion of the grouping decreases after the merge and the process is reiterated until not further merge can decrease the criterion.

The algorithm relies on the $O(I)$ marginal counts, which require one $O(n)$ scan of the data set. However, we express the complexity of the algorithm in terms of n (rather than in terms of I), since the number of categorical values can reach $O(n)$ in the worst case. 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 (Boullé, 2004a). The algorithm is mainly based on 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 Δ value. These Δ values 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 Δ values 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 Δ values related to all the possible merges of 2 values: $O(n^2)$
 - 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 Δ values of the remaining group merges adjacent to the best merge: $O(n)$
 - Update the sorted list of merges: $O(n \log(n))$

In the case of two classes, the time complexity of the greedy algorithm can be optimized down to $O(n \log(n))$ owing to conjecture 1.

2.4.2 Preprocessing

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, the initial categorical values can be preprocessed into a new set of values of cardinality $I' \leq \sqrt{n}$.

A first straightforward preprocessing step is to merge pure values having the same class. This step is compatible with the optimal solution (see Theorem 2).

A second preprocessing step consists in building J groupings for each "one class against the others" sub-problems. This require $O(J n \log(n))$ time. The subparts of the groups shared by all the J groupings can easily be identified and represent very good candidate subgroups of the global grouping problem. The number of these subparts is usually far below the number of initial categorical values and helps achieving a reduced sized set of preprocessed values.

A third preprocessing step can be applied when the number of remaining preprocessed values is beyond \sqrt{n} . The values can be sorted by decreasing frequencies and the exceeding infrequent values can be unconditionally grouped into J groups according to their majority class. This last step is mandatory to control the computational complexity of the algorithm. However, experiments show that this last step is rarely activated in practice.

2.4.3 Post-Optimizations

The greedy heuristic may fall in a local optimum, so that time efficient post-optimizations are potentially useful to improve the quality of the solution. Since the evaluation criterion is Bayes optimal, spending more computation time is meaningful.

A first post-optimization step consists in forcing the merges between groups until a single terminal group is obtained and to keep the best encountered grouping. This helps escaping local optima and requires $O(n \log(n))$ computation time. Furthermore, in the case of noisy attribute where the optimal grouping consists of a single group, this heuristic guarantees to find the optimal solution.

A second post-optimization step consists in evaluating every move of a categorical value from one group to another. The best moves are performed as long as they improve the evaluation criterion. This process is similar to the K-means algorithm where each value is attracted by its closest group. It converges very quickly, although this cannot be proved theoretically.

A third post-optimization step is a look-ahead optimization. The best merge between groups is simulated and post-optimized using the second step algorithm. The merge is performed in case of improvement. This algorithm looks similar to the initial greedy merge algorithm, except that it starts from a very good solution and incorporates an enhanced post-optimization. Thus, this additional post-optimization is usually triggered for only one or two extra merges.

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 on artificial and real data sets. Finally, we present the impact of grouping as a preprocessing step to the Naïve Bayes classifier.

3.1 Presentation

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 (Boullé, 2004b), where each grouping method is considered as an elementary inductive method which predicts the distribution of the class values in each learned groups.

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 class conditional information contained in the attribute while decreasing the number of values. In the experiments, we collect both the number of groups and the predictive quality of the grouping. The number of groups is easy to calculate. The quality of the estimation of the class conditional information hold by each group is more difficult to evaluate.

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. In the case of categorical attributes, we have the unique opportunity of observing the class conditional distribution on the test data set: for each categorical value in the test data set, the observed distribution of the class values can be estimated by counting. The grouping methods allow to induce the class conditional distribution from the train data set: for each learnt group on the train data set, the learnt distribution of the class values can be estimated by counting. The objective of grouping is to minimize the distance between the learnt distribution and the observed distribution. This distance can be evaluated owing to a divergence measure, such as the Kullback-Leibler divergence, the chi-square divergence or the Hellinger coefficient. In our experiments, we choose to evaluate this distance using the Kullback-Leibler divergence (Kullback, 1968).

The MODL grouping methods exploits a space of class conditional distribution models (the SGM models) and searches the most probable model given the train data. It is noteworthy that no loss function is optimized in this approach: neither the classification accuracy is optimized nor the Kullback-Leibler divergence (which would require to divide the train data set into train data and validation data). The MODL method exploits all the available train data to build its grouping model.

The evaluation conducted in the experiments focuses on the quality of the groupings (size and Kullback-Leibler divergence criterions) as a preprocessing method for data mining. It is interesting to examine whether optimizing the posterior probability of a grouping model (on the basis on the three-stage prior) leads to high-quality groupings.

3.2 The Evaluation Protocol

The predictive quality of the groupings is evaluated owing to the Kullback-Leibler divergence (Kullback, 1968) applied to compare the distribution of the class values estimated from the train data set with the distribution of the class values observed on the test data set.

Let n' be the number of instances, n'_i be the number of instances for value i and n'_{ij} the number of instances for value i and class j on the test data set.

For a given categorical value i , let p_{ij} be the probability of the j^{th} class value estimated on the train data set (on the basis of the group containing the categorical value), and q_{ij} be the probability of the j^{th} class value observed on the test data set (using directly the categorical value). The q_{ij} probabilities are estimated with the Laplace's estimator in order to deal with zero values. We get

$$p_{ij} = \frac{n_{k(i)j}}{n_{k(i)}} \quad \text{and} \quad q_{ij} = \frac{n'_{ij} + 1}{n'_i + J}.$$

The mean of the Kullback-Leibler divergence on the test data set is

$$D(p||q) = \sum_{i=1}^I \frac{n'_i}{n'} \sum_{j=1}^J p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

In a first experiment, we compare the behavior of the evaluated grouping method on synthetic data sets, where the ideal grouping pattern is known in advance. In the second experiments, we

use real data sets to compare the grouping methods considered as univariate classifiers. In a third experiment, we use the same data sets to evaluate the results of Naïve Bayes classifiers using the grouping methods to preprocess the categorical attributes. In this last experiment, the results are evaluated using the classification accuracy both on train data sets and on test data sets.

Data Set	Continuous Attributes	Categorical Attributes	Size	Class Values	Majority Accuracy
Adult	7	8	48842	2	76.07
Australian	6	8	690	2	55.51
Breast	10	0	699	2	65.52
Crx	6	9	690	2	55.51
Heart	10	3	270	2	55.56
HorseColic	7	20	368	2	63.04
Ionosphere	34	0	351	2	64.10
Mushroom	0	22	8416	2	53.33
TicTacToe	0	9	958	2	65.34
Vehicle	18	0	846	4	25.77
Waveform	40	0	5000	3	33.84
Wine	13	0	178	3	39.89

Table 1. Data sets

We gathered 12 data sets from U.C. Irvine repository (Blake and Merz, 1998), each data set has at least a few tenths of instances for each class value and some categorical attributes with more than two values. In order to increase the number of categorical attributes candidate for grouping, the continuous attributes have been discretized in a preprocessing step with a 10 equal-width unsupervised discretization. The 12 data sets comprising 230 attributes are described in Table 1; the last column corresponds to the accuracy of the majority class.

The categorical attributes in these data sets hold less than 10 values on average (from an average 3 values in the TicTacToe attributes to about 20 values in the HorseColic attributes). In order to perform more discriminating experiments, we use a second collection of data sets containing all the cross-products of the attributes. In this "bivariate" benchmark, the 12 data sets contain 2614 categorical attributes holding 55 values on average.

3.3 The Evaluated Methods

The grouping methods studied in the comparison are:

- MODL, the method described in this paper,
- Khiops (Boulle, 2004b),
- BIC (Ritschard, 2003),
- CHAID (Kass, 1980),
- Tschuprow (Ritschard et al., 2001),
- Gain Ratio (Quinlan, 1993).

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 method is based on a Bayesian approach and incorporates preprocessing and post-optimization algorithms. The Khiops, CHAID, Tschuprow and BIC methods use the chi-square statistics in different manner. The Gain Ratio method is based on entropy.

The CHAID method is the grouping method used in the CHAID decision tree classifier (Kass, 1980). It applies the chi-square criterion locally to two categorical values in the

contingency table, and iteratively merges the values as long as they are statistically similar. The significance level is set to 0.95 for the chi-square threshold. 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. 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. The BIC method is based on the deviance G^2 statistics, which is a chi-square statistics. It exploits a Bayesian information criterion (Schwarz, 1978) to select the best compromise model between fit and complexity. The Gain Ratio method is the methods used in the C4.5 decision tree classifier (Quinlan, 1993). The gain ratio criterion attempts to find a trade-off between the information on the class values (information gain) and the complexity of the partition (the split information) by dividing the two quantities.

We have re-implemented these alternative grouping approaches in order to eliminate any variance resulting from different cross-validation splits.

3.4 The Artificial Data Sets Experiments

Using artificial data sets allows controlling the distribution of the explanatory values and of the class values. The evaluation of the groupings learned on train data sets can thus be optimal, without requiring any test data set. In the case of grouping, an artificial data set containing one categorical attribute and one class attribute is completely specified by the following parameters:

I : number of categorical values,

J : number of classes,

p_i , $1 \leq i \leq I$: probability distribution of the categorical values,

p_{ji} , $1 \leq j \leq J$, $1 \leq i \leq I$: probability distribution of the class values conditionally to the categorical values.

3.4.1 The Noise Pattern

The purpose of this experiment is to compare the robustness of the grouping methods. The *noise pattern* data set consists of an explanatory categorical attribute independent from the class attribute. The explanatory attribute is uniformly distributed ($p_i=1/I$) and the class attribute consists of two equidistributed class values ($p_{ji}=1/2$). We use randomly generated train samples of size 1000 and perform the experiment 1000 times, for different numbers of categorical values. In the case of independence, the optimal number of groups is 1. In Figure 2, we report the mean of the number of unnecessary groups ($K-1$) and of the Kullback-Leibler divergence between the estimated class distribution and the true class distribution.

The results demonstrate the effectiveness of the Kullback-Leibler divergence to evaluate the quality of a grouping. In the case of attribute independence, the classification accuracy is uniformly equal to 50% whatever the number of groups: it is useless as an evaluation criterion. The most comprehensible grouping consists of a single group whereas the worst one is to build one group per value. The Kullback-Leibler divergence is able to exploit this by a better evaluation of the true class distribution in the most frequent groups. Thus, building separate groups that could be merged leads to a poorer estimation of the class distribution.

The CHAID method creates more and more groups when the number of categorical values increases. This translates by a quickly decreasing quality of estimation of the class distribution, as pictured in Figure 2.

The BIC method performs better than the CHAID method when the number of categorical values is small. When this number of values increases, the BIC and CHAID methods perform similarly.

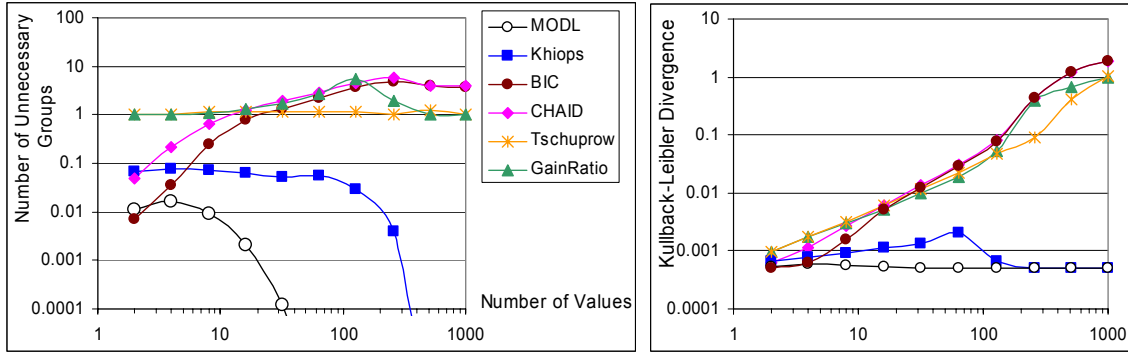


Figure 2. Mean of the number of unnecessary groups ($K-1$) and of the Kullback-Leibler divergence of the groupings of an explanatory attribute independent from the class attribute

The GainRatio method is constrained to produce at least 2 groups. It overfits the train data as soon as the number of categorical values exceeds a few tens.

The Tschuprow method is also constrained to produce at least 2 groups. It builds almost systematically exactly two groups. A closer look at the Tschuprow criterion shows that the criterion can reach its theoretical bound only when the contingency table is square, meaning that the number of groups is exactly equal to the number of class values. Additional experiments with different numbers of class values (not reported in this paper) confirm this bias. Although the number of groups is constant under varying the number of categorical values, the estimation of the class distribution worsens with higher number of values. This is explained since less frequent values lead to a less reliable estimation of the class probabilities.

The Khiops method is designed to build one single group with probability 0.95, when the explanatory attribute and the class attribute are independent. This is confirmed by the experiment up to about 100 categorical values. Above this number of values, the Khiops method systematically builds one group for the reason that it unconditionally groups the least frequent values in a preprocessing step. The objective of this preprocessing step is 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.

The MODL method builds one single group almost always, and the proportion of multi-groups partitions decreases sharply with the number of categorical values. The experiments have been performed 100000 times for the MODL method. Above 50 values, no grouping (out of 100000) contains more than one group.

3.4.2 The Mixture Pattern

The objective of this experiment is to evaluate the sensibility of the grouping methods. The *mixture pattern* data set consists of an explanatory categorical attribute distributed according to several mixtures of class values. The explanatory attribute is uniformly distributed ($p_i=1/I$) and the class attribute consists of four equidistributed class values ($p_{j/i}=1/4$). We designed 8 artificial groups corresponding to 8 different mixtures of class values. In each group, one of the class values is the majority class (with probability 0.50 or 0.75) and the other three classes are equidistributed, as pictured in Figure 3.

We use randomly generated train samples of size 10000 and perform the experiment 1000 times, for different numbers of categorical values. Due to the quadratic complexity of the algorithms, the experiment was conducted up to 2000 categorical values, except for the MODL algorithm that has a super-linear complexity. In Figure 4, we report the mean of the number of groups and of the Kullback-Leibler divergence between the estimated class distribution and the true class distribution.

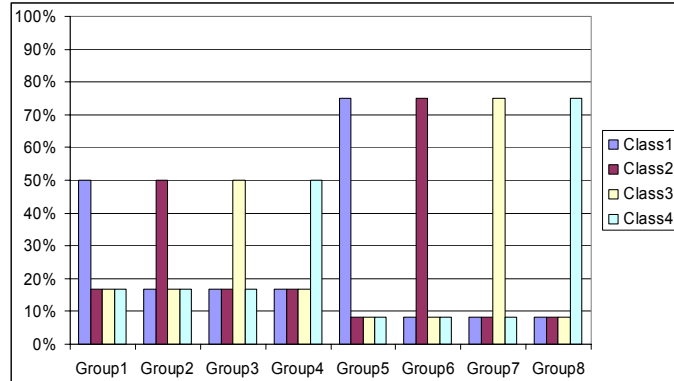


Figure 3. Class distribution in 8 artificial group and 4 class values; the categorical values are uniformly distributed on the 8 groups

The CHAID method overfits the training data by creating too many groups, even more than in the case of independence since the number of class values is now 4. The BIC method manages to find the optimal number of groups when the number of categorical values is below 100. Above this threshold, it overfits the training data and exhibits a behavior similar to that of the CHAID method. The behavior of the GainRatio method is unexpected. For small numbers of categorical values, it builds a constant number of groups equal to the number of class values, and beyond one hundred values, the number of groups raises sharply. The Tschuprow method is so strongly biased that it always produces exactly 4 groups. The Khiops method benefits from its robustness and correctly identifies the 8 artificial groups as long as the frequencies of the categorical values are sufficient. Beyond about 400 values, the minimum frequency constraint of the Khiops algorithms become active and the number of groups falls down to 1. The MODL method almost always builds optimal groupings with the correct number of groups. When the number of categorical values becomes large (about 500, *i.e.* an average frequency of 40 per value), there is a transition in the behavior of the algorithm, that produces only 4 groups instead of 8. The frequency of the categorical values is no longer sufficient to discriminate 8 types of class distributions. When the number of class values increases again (beyond 2000, *i.e.* an average frequency of 5 per value), there is a second transition and the MODL method builds one single group.

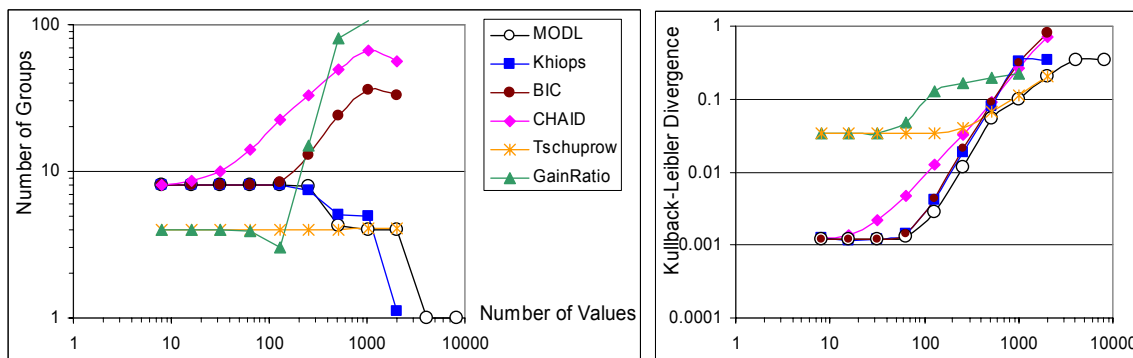


Figure 4. Mean of the number of groups and of the Kullback-Leibler divergence of the groupings of an explanatory attribute distributed in 8 mixtures of class values

To summarize, the noise and mixture pattern experiments are a convenient way to characterize each grouping methods with their bias, robustness, sensibility and limits. The experiments show the interest of using the Kullback-Leibler divergence to evaluate the quality of

the groupings. This criterion looks well suited to evaluate the groupings in real data sets, where the true class distribution is unknown.

The CHAID method exhibits an overfitting behavior, which decays when the number of categorical values or class values increases. The BIC method finds the correct number of groups when the categorical values are not too numerous and then overfits the training data. The Tschuprow is strongly biased in favor of numbers of groups equal to numbers of classes. The GainRatio exhibits a varying biased and overfitting behavior according to the distribution of the train data. The Khiops method is robust but suffers from a lack of sensibility when the categorical values are too numerous, due to its minimum frequency constraint.

The MODL method builds groupings that are optimum, the most probable groupings given the train data. It is the only evaluated method that retrieves the exact number of groups both in case of noise data and of informative data.

3.5 The Real Data Sets Experiments

The goal of this 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 all the attributes of the UCI data sets presented in Table 1, using a stratified tenfold cross-validation. As the purpose of the experiments is to evaluate the grouping methods according to the way they compromise between the quality and the size of the groupings, we also added three basic grouping methods for comparison reasons. The first method named NoGrouping builds one group per categorical value: it is the least biased method for estimating the distribution of the class values at the expense of the highest number of groups. The second method called ExhaustiveCHAID (SPSS, 2001) is a version of the CHAID method that merges similar pairs continuously until only a single pair remains. We added a similar method ExhaustiveMODL, which allows comparing the two methods when they are constrained to build the same number of groups.

During the experiments, we collect the number of groups and the Kulback-Leibler divergence between the class distribution estimated on train data sets and the class distribution observed on test data sets. For each grouping method, this represents 2300 measures for the univariate analysis (230 attributes) and 26140 measures for the bivariate analysis (2614 pairs of attributes). All these results are summarized across the attributes of the data sets owing to means, in order to provide a gross estimation of the relative performances of the methods. We report the mean of the number of groups and of the Kullback-Leibler divergence for the univariate and bivariate analysis in Figure 5. For the Kullback-Leibler divergence, we use geometric means normalized by the result of the NoGrouping method in order to focus on the ratios of predictive performance between tested methods. The gray line highlights the Pareto curve of the results obtained by the grouping methods.

As expected, the NoGrouping method obtains the best results in term of predictive quality, at the expense of the worst number of groups. However, in the univariate analysis, the Khiops, BIC, CHAID and MODL methods reach almost the same quality with far less groups. The Tschuprow method is hampered by its bias in favor of number of groups equal to the number of class values, so that its performance are not better that those of groups equal to the number of class values, ExhaustiveCHAID and ExhaustiveMODL. The GainRatio method is dominated by the other methods. The bivariate analysis is much more selective. The results follow the same trend with sharper differences between the methods. The Khiops method underfits the data because of its minimum frequency constraint, while the CHAID method suffers from its lack of overfitting control by producing too many groups and degrading its predictive performance. Although its criterion incorporates a complexity penalty, the BIC method builds too many groups and overfits the data.

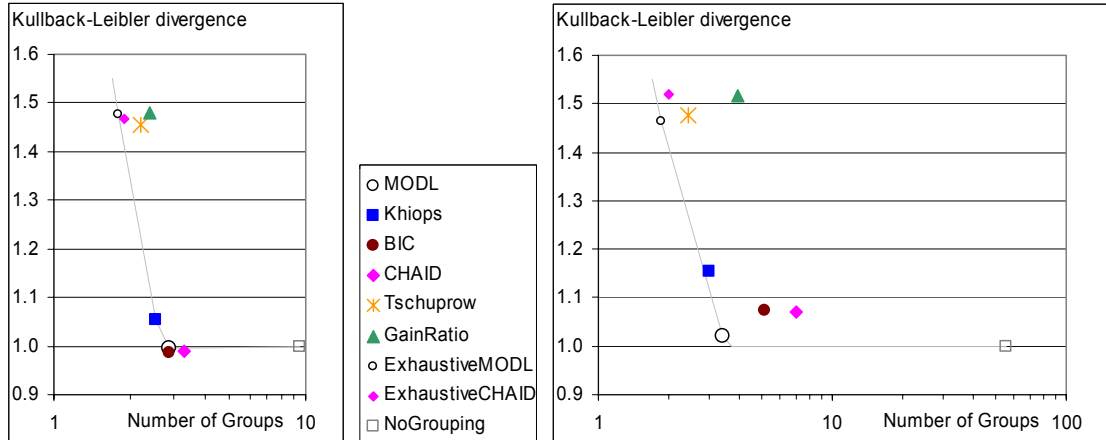


Figure 5. Mean of the number of groups and of the Kullback-Leibler divergence of the groupings performed on the UCI data sets, in univariate analysis (on the left) and bivariate analysis (on the right)

The MODL method gets the lowest number of group without discarding the predictive quality. It manages to reduce the number of categorical values by one order of magnitude while keeping the best estimate of the class conditional probability.

3.6 Impact of Groupings on the Naïve Bayes Classifier

The aim of this experiment is to evaluate the impact of grouping methods on the Naïve Bayes classifier. The Naïve Bayes classifier (Langley et al., 1992) assigns the most probable class value given the explanatory attributes values, assuming conditional independence between the attributes for each class value. The Naïve Bayes classifier is a very simple technique that does not require any parameter setting, which removes the bias due to parameter tuning. It performs surprisingly well on many data sets (Dougherty et al., 1995; Hand and Yu, 2001) and is very robust. High quality grouping tend to increase the frequency in each group and to decrease the variance in the estimation of the conditional class density. It is interesting to examine whether this can benefit to classifiers, even to the Naïve Bayes classifier which is particularly unsophisticated.

The evaluation of probabilities for numeric attributes owing to discretization has already been discussed in the literature (Dougherty et al., 1995; Hsu et al, 2003; Yang and Webb, 2003). Experiments have shown that even a simple Equal Width discretization with 10 bins brings superior performances compared to the assumption using the Gaussian distribution. On the opposite, the probabilities for categorical attribute are estimated using the Laplace's estimator directly on the categorical values, without any preprocessing such as grouping. It is interesting to study whether grouping could produce a more robust estimation of the class distributions and enhance the performance of the Naïve Bayes classifier. The experiment is performed on the 12 data sets presented in Table 1, using the univariate (all categorical attributes) and bivariate (all pairwise interactions of categorical attributes) sets of data sets. A Student's test at the 5% confidence level is performed between the MODL grouping method and the other methods to determine whether the differences of performance are significant. According to (Dietterich, 1998), the McNemar's test is more reliable, but it does not assess the effect of varying the training set. However, these statistical tests "must be viewed as approximate, heuristic tests, rather than as rigorously correct statistical methods" (Dietterich, 1998). Table 2 reports a summary of the test accuracy and robustness results, using the mean of the data set results, the average rank of each method and the number of significant wins and losses of the MODL method.

	Test accuracy						Robustness (Test acc. / Train acc.)					
	Univariate data sets			Bivariate data sets			Univariate data sets			Bivariate data sets		
	Mean	Rank	Wins	Mean	Rank	Wins	Mean	Rank	Wins	Mean	Rank	Wins
MODL	84.8%	2.6		85.8%	2.8		98.6%	3.0		96.8%	4.2	
Khiops	84.1%	3.4	3/1	83.7%	4.7	4/0	98.6%	3.8	1/0	97.2%	2.5	0/3
BIC	83.2%	4.2	4/0	84.9%	3.2	3/0	96.1%	5.2	3/0	93.9%	6.7	7/0
CHAID	83.3%	3.7	2/0	84.7%	4.0	2/1	96.1%	5.2	1/0	93.6%	7.3	7/0
Tschuprow	82.8%	6.4	7/0	83.3%	6.6	6/0	96.6%	5.2	1/0	93.8%	5.5	3/0
GainRatio	81.5%	6.1	5/1	82.9%	5.8	4/0	95.4%	6.0	1/0	93.0%	5.7	5/0
ExMODL	83.4%	5.2	5/0	84.6%	4.7	4/0	98.5%	3.8	0/0	97.1%	2.7	0/2
ExCHAID	81.9%	7.4	7/0	83.2%	5.7	6/0	96.2%	5.8	2/0	94.6%	3.5	2/2
NoGrouping	84.0%	4.1	4/0	84.6%	5.1	4/0	97.3%	6.4	3/0	94.0%	6.2	6/0

Table 2. Summary of the test accuracy and robustness results (mean, average rank and number of wins/losses) of the Naïve Bayes classifier on the UCI data sets, in univariate analysis and bivariate analysis

The results look consistent on the three indicators and show that the MODL method dominates the other methods on the test accuracy criterion. The mean results are pictured in Figure 6 with the classification accuracy reported both on train and test data sets, in order to visualize the train and test accuracy of the methods in a two criteria-analysis. The thick gray line on the diagonal represents the asymptotic best achievable robustness of the methods. The thin gray line highlights the Pareto curve in the two-criteria analysis between robustness and test accuracy.

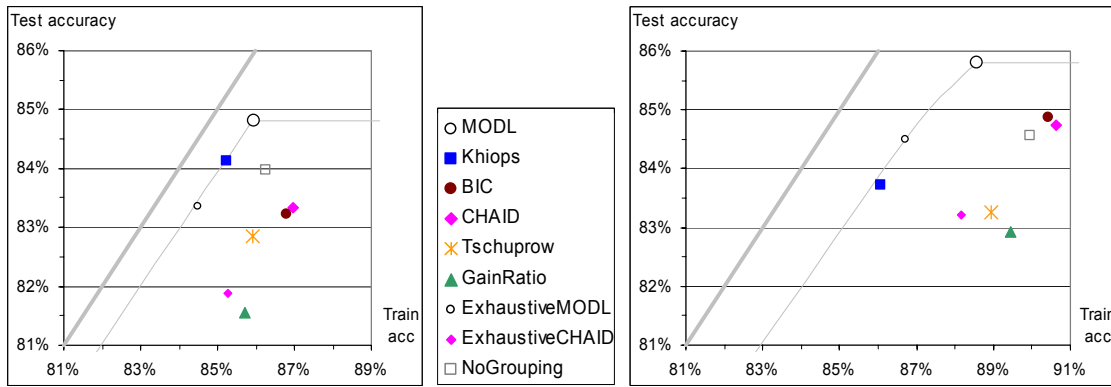


Figure 6. Mean of the train accuracy and test accuracy of the Naïve Bayes classifier on the UCI data sets, in univariate analysis (on the left) and bivariate analysis (on the right)

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 MODL method clearly dominates all the other methods, owing to the quality of its groupings. The resulting Naïve Bayes classifier is both the more robust one (together with Khiops) and the more accurate on test data sets. Another important aspect learnt from this experiment is the overall gain in test accuracy when the pairs of attributes are used. The bivariate analysis 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. From univariate to bivariate analysis, the MODL method achieves an increase in test accuracy about twice and a decay in robustness approximately half that of the reference NoGrouping method.

Compared to the NoGrouping method, the MODL method is the only evaluated grouping method that improves the test accuracy of the Naïve Bayes classifier. The most noticeable effect of using the MODL method is a drastic improvement of the robustness.

4 Conclusion

When categorical attributes contain few values, typically less than 10 values, using a grouping method is not required. As the number of values increases (which is common in marketing applications), preprocessing the categorical attributes becomes attractive in order to improve the performance of classifiers. The issue of grouping methods is to reduce the number of groups of values while maintaining the conditional class information.

The MODL grouping method exploits the 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. An optimization heuristics including preprocessing and post-optimizations is proposed in this paper to optimize the grouping with super-linear time complexity. This algorithm manages to efficiently find high quality groupings.

Extensive evaluations both on real and synthetic data indicate notable performances for the MODL method. It is time efficient and does not require any parameter setting. It builds groupings that are both robust and accurate. The more valuable characteristic of the MODL method is probably the understandability of the groupings. Although understandability is hard to evaluate, the method is theoretically founded to produce correct explanations of the explanatory categorical attributes on the basis of the partition of their values, and even the most probable "grouping based" explanation given the train data.

Acknowledgments

I am grateful to the editor and the anonymous reviewers for their useful comments.

Appendix A

In this appendix, we first present the combinatorial formula used to evaluate the numbers of partition of a set and second provide the proof of the theorems introduced in the paper.

A.1 Partition numbers

The number of ways a set of n elements can be partitioned into nonempty subsets is called a Bell number and denoted B_n .

The number of ways of partitioning a set of n elements into k nonempty subsets is called a Stirling number of the second kind and denoted $S(n, k)$.

The Bell numbers can be defined by the sum

$$B_n = \sum_{k=1}^n S(n, k) .$$

Let $B(n, k)$ be the number of partition of a set of n elements into at most k parts. This number, that we choose to call a generalized Bell number, can be defined by the sum

$$B(n, k) = \sum_{i=1}^k S(n, i) .$$

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:

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

Proof:

The prior probability of a grouping model M can be defined by the prior probability of the parameters of the model.

Let us introduce some notations:

- $p(K)$: prior probability of the number of groups K ,
- $p(\{k(i)\})$: prior probability of a partition (defined by $\{k(i)\}$) of the categorical values into K groups,
- $p(\{n_{kj}\})$: prior probability of the set of parameters $\{n_{11}, \dots, n_{kj}, \dots, n_{KJ}\}$,
- $p(\{n_{kj}\}_k)$: prior probability of the set of parameters $\{n_{k1}, \dots, n_{kJ}\}$.

The objective is to find the grouping model M that maximizes the probability $p(M | D)$ for a given train data set D . Using Bayes formula and since the probability $p(D)$ is constant under varying the model, this is equivalent to maximize $p(M)p(D|M)$.

Let us first focus on the prior probability $p(M)$ of the model. We have

$$\begin{aligned} p(M) &= p(K, \{k(i)\}, \{n_{kj}\}) \\ &= p(K) p(\{k(i)\} | K) p(\{n_{kj}\} | K, \{k(i)\}). \end{aligned}$$

The first hypothesis of the three-stage prior is that the number of groups is uniformly distributed between 1 and I . Thus we get

$$p(K) = \frac{1}{I}.$$

The second hypothesis is that all the partition of the categorical values into at most K groups are equiprobable for a given K . Computing the probability of one set of groups turns into the combinatorial evaluation of the number of possible group sets. This number is equal to the generalized Bell number $B(I, K)$. Thus we obtain

$$p(\{k(i)\} | K) = \frac{1}{B(I, K)}.$$

The last term to evaluate can be rewritten as a product using the hypothesis of independence of the distributions of the class values between the groups. We have

$$\begin{aligned} p(\{n_{kj}\} | K, \{k(i)\}) &= p(\{n_{kj}\}_1, \{n_{kj}\}_2, \dots, \{n_{kj}\}_K | K, \{k(i)\}) \\ &= \prod_{k=1}^K p(\{n_{kj}\}_k | K, \{k(i)\}) \\ &= \prod_{k=1}^K p(\{n_{kj}\}_k | n_k). \end{aligned}$$

The frequencies per group n_k derive from the frequencies per categorical values n_i for a given partition of the values into groups.

For a given group k with size n_k , all the distributions of the class values are equiprobable. Computing the probability of one distribution is a combinatorial problem, which solution is

$$p(\{n_{kj}\}_k | n_k) = \frac{1}{C_{n_k+J-1}^{J-1}}.$$

Thus,

$$p(\{n_{kj}\} | K, \{k(i)\}) = \prod_{k=1}^K \frac{1}{C_{n_k+J-1}^{J-1}}.$$

The prior probability of the model is then

$$p(M) = \frac{1}{I} \frac{1}{B(I, K)} \prod_{k=1}^K \frac{1}{C_{n_k+J-1}^{J-1}}.$$

Let us now evaluate the probability of getting the train data set D for a given model M . We first divide the data set D into K subsets D_k of size n_k corresponding to the K groups. Using again the independence assumption between the groups, we obtain

$$\begin{aligned} p(D | M) &= p(D | K, \{k(i)\}, \{n_{kj}\}) \\ &= p(D_1, D_2, \dots, D_K | K, \{k(i)\}, \{n_{kj}\}) \\ &= \prod_{k=1}^K p(D_k | K, \{k(i)\}, \{n_{kj}\}) \\ &= \prod_{k=1}^K \frac{1}{(n_k! / n_{k,1}! n_{k,2}! \dots n_{k,J}!)} \end{aligned}$$

as evaluating the probability of a subset D_k under uniform prior turns out to be a multinomial problem.

Taking the negative log of the probabilities, the maximization problem turns into the minimization of the claimed criterion

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

Theorem 2: In a Bayes optimal SGM model distributed according to the three-stage prior, two pure categorical values having the same class are necessary in the same group.

Proof:

Let a and b be the two pure values related to the same class (indexed as class 1 for convenience reasons). Let us assume that, contrary to the claim, the two values are separated into two groups $A1$ and $B1$.

We construct two new groups $A0$ and $B2$ by moving the pure value a from $A1$ to $B1$. The cost variation $\Delta Cost1$ of the grouping is

$$\begin{aligned} \Delta Cost1 &= \Delta PartitionCost1 \\ &+ \log((n_{A0} + J - 1)! / n_{A0}!(J - 1)!) + \log(n_{A0}! / n_{A0,1}! n_{A0,2}! \dots n_{A0,J}!) \\ &+ \log((n_{B2} + J - 1)! / n_{B2}!(J - 1)!) + \log(n_{B2}! / n_{B2,1}! n_{B2,2}! \dots n_{B2,J}!) \\ &- \log((n_{A1} + J - 1)! / n_{A1}!(J - 1)!) - \log(n_{A1}! / n_{A1,1}! n_{A1,2}! \dots n_{A1,J}!) \\ &- \log((n_{B1} + J - 1)! / n_{B1}!(J - 1)!) - \log(n_{B1}! / n_{B1,1}! n_{B1,2}! \dots n_{B1,J}!). \end{aligned}$$

If the $A1$ group contains only the a value, moving a from $A1$ to $B1$ results in a decreased number of groups, with a related variation of partition cost

$$\Delta PartitionCost1 = \log(B(I, K - 1)) - \log(B(I, K)),$$

which is negative. In the opposite case, the number of groups remains the same and the resulting variation of partition cost is zero.

The frequencies are the same for each class except for class 1, thus

$$\begin{aligned} \Delta Cost1 - \Delta PartitionCost1 &= \log\left(\frac{(n_{A1} - n_a + J - 1)!}{(n_{A1} + J - 1)!}\right) \\ &\quad + \log\left(\frac{(n_{B1} + n_a + J - 1)!}{(n_{B1} + J - 1)!}\right) \\ &\quad + \log\left(\frac{n_{A1,1}!}{(n_{A1,1} - n_{a,1})!}\right) \\ &\quad + \log\left(\frac{n_{B1,1}!}{(n_{B1,1} + n_{a,1})!}\right), \\ \Delta Cost1 - \Delta PartitionCost1 &= \log\left(\frac{\prod_{n=0}^{n_a-1} (n_{A1,1} - n)}{(n_{A1} - n + J - 1)}\right) \\ &\quad - \log\left(\frac{\prod_{n=1}^{n_a} (n_{B1,1} + n)}{(n_{B1} + n + J - 1)}\right). \end{aligned}$$

Similarly, we construct two groups $A2$ and $B0$ by moving pure value b from $B1$ to $A1$. This time, the cost variation $\Delta Cost2$ of the grouping is

$$\begin{aligned} \Delta Cost2 - \Delta PartitionCost2 &= \log\left(\frac{\prod_{n=0}^{n_b-1} (n_{B1,1} - n)}{(n_{B1} - n + J - 1)}\right) \\ &\quad - \log\left(\frac{\prod_{n=1}^{n_b} (n_{A1,1} + n)}{(n_{A1} + n + J - 1)}\right). \end{aligned}$$

Let us assume that

$$n_{A1,1}/(n_{A1} + J - 1) \leq n_{B1,1}/(n_{B1} + J - 1).$$

Using the property

$$0 \leq z \leq x < y \Rightarrow (x - z)/(y - z) \leq x/y \leq (x + z)/(y + z),$$

we obtain

$$\begin{aligned} \prod_{n=0}^{n_a-1} (n_{A1,1} - n)/(n_{A1} - n + J - 1) &\leq (n_{A1,1}/n_{A1} + J - 1)^{n_a} \\ &\leq (n_{B1,1}/n_{B1} + J - 1)^{n_a} \leq \prod_{n=1}^{n_a} (n_{B1,1} + n)/(n_{B1} + n + J - 1). \end{aligned}$$

Thus, we get $\Delta Cost1 - \Delta PartitionCost1 \leq 0$.

On the opposite, let us assume that

$$n_{A1,1}/(n_{A1} + J - 1) \geq n_{B1,1}/(n_{B1} + J - 1).$$

This time, we obtain $\Delta Cost2 - \Delta PartitionCost2 \leq 0$.

Since the variations of partition costs are always non-negative, at least one of the two cost variations $\Delta Cost1$ or $\Delta Cost2$ is negative and the initial grouping could not be optimal. As this is contradictory with the initial assumption, the claim follows.

Remark:

If the partition costs are evaluated using the Stirling numbers of the second kind instead of the generalized Bell numbers, this theorem is no longer true, since decreasing the number of groups can result in an increase of the partition cost (for example, $S(I, I-1) = I(I-1)/2$ and $S(I, I) = 1$).

Theorem 3: In a SGM model distributed according to the three-stage prior and in the case of two classes, the Bayes optimal grouping model consists of a single group when each instance has a different categorical value.

Proof:

Since each instance has a different categorical value, all the categorical values are pure values associated with one among the J class values. According to Theorem 2, the values having the same class values are necessary in the same group. The optimal grouping contains at most J groups.

Let A and B be two groups and $A \cup B$ the group obtained by merging A and B .

Let n_A , n_B and $n_{A \cup B}$ be the frequencies of these groups, and let $n_{A,j}$, $n_{B,j}$ and $n_{A \cup B,j}$ be the frequencies per class value in these groups.

When the two groups are merged, the number of groups decreases from K to $K-1$ with the variation of partition cost $\Delta PartitionCost = \log(B(I, K-1)) - \log(B(I, K))$.

The total variation of the grouping cost is

$$\begin{aligned} \Delta Cost &= \Delta PartitionCost \\ &+ \left(\log \left(C_{n_{A \cup B} + J - 1}^{J-1} \right) + \log \left(n_{A \cup B}! / n_{A \cup B,1}! n_{A \cup B,2}! \dots n_{A \cup B,J}! \right) \right) \\ &- \left(\log \left(C_{n_A + J - 1}^{J-1} \right) + \log \left(n_A! / n_{A,1}! n_{A,2}! \dots n_{A,J}! \right) \right) \\ &- \left(\log \left(C_{n_B + J - 1}^{J-1} \right) + \log \left(n_B! / n_{B,1}! n_{B,2}! \dots n_{B,J}! \right) \right), \\ \Delta Cost &= \Delta PartitionCost \\ &+ \log \left((n_{A \cup B} + J - 1)! (J - 1)! / (n_A + J - 1)! (n_B + J - 1)! \right) \\ &- \sum_{j=1}^J \log \left(C_{n_{A \cup B,j}}^{n_{A,j}} \right). \end{aligned}$$

Since each class is fully contained either in group A or B , we obtain

$$\Delta Cost = \Delta PartitionCost + \log \left((n_{A \cup B} + J - 1)! (J - 1)! / (n_A + J - 1)! (n_B + J - 1)! \right).$$

We are in the case of 2 class values and thus have $J = 2$, $n_{A \cup B} = n$, $K = 2$.

$$\begin{aligned} \Delta Cost &= \log(B(n, 1)) - \log(B(n, 2)) + \log \left((n+1)! / (n_A+1)! (n_B+1)! \right) \\ &= -\log(2^{n-1}) + \log \left(C_{n+2}^{n_A+1} \right) - \log(n+2). \end{aligned}$$

Since we have $C_n^k = C_{n-1}^{k-1} + C_{n-1}^k \leq 2^{n-1}$ for $n > 1$ and $k > 1$ (with a strict inequality when $n > 2$), we finally obtain

$$\begin{aligned} \Delta Cost &< -\log(2^{n-1}) + \log(2^{n+1}) - \log(n+2) \\ &< \log(4/(n+2)) \\ &< 0. \end{aligned}$$

The claim follows.

Remark:

If the partition costs are evaluated using the Stirling numbers of the second kind instead of the generalized Bell numbers, this theorem is no longer true. In particular, when the class values are equi-distributed, the grouping cost in the case of a single group ($\log(n) + \log(n+1) + \log(C_n^{n/2})$) is higher than the grouping cost in the case of one group per categorical value ($\log(n) + n \log(2)$).

References

M. Asseraf. Metric on decision trees and optimal partition problem. *International Conference on Human System Learning, Proceedings of CAPS'3*, Paris, 2000.

- N. C. Berckman. Value grouping for binary decision trees. Technical Report, Computer Science Department – University of Massachusetts, 1995.
- J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, New York, 1994.
- C. L. Blake and C. J. Merz. UCI Repository of machine learning databases Web URL <http://www.ics.uci.edu/~mllearn/MLRepository.html>. Irvine, CA: University of California, Department of Information and Computer Science, 1998.
- M. Boullé. Khiops: a Statistical Discretization Method of Continuous Attributes. *Machine Learning*, 55(1):53-69, 2004a.
- M. Boullé. 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: 173-182, 2004b.
- M. Boullé. A Bayesian Approach for Supervised Discretization. *Data Mining V*, Eds Zanasi, Ebecken, Brebbia, WIT Press, pp 199-208, 2004c.
- L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. *Classification and Regression Trees*. California: Wadsworth International, 1984.
- B. Cestnik, I. Kononenko and I. Bratko. ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In Bratko and Lavrac (Eds.), *Progress in Machine Learning*. Wilmslow, UK: Sigma Press, 1987.
- P. A. Chou. Optimal Partitioning for Classification and Regression Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):340-354, 1991.
- T. G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Methods. *Neural Computation*, 10(7), 1998.
- J. Dougherty, R. Kohavi and M. Sahami. Supervised and Unsupervised Discretization of Continuous Features. *Proceedings of the Twelfth International Conference on Machine Learning*. Los Altos, CA: Morgan Kaufmann, pp 194-202, 1995.
- T. Fulton, S. Kasif and S. Salzberg. Efficient algorithms for finding multi-way splits for decision trees. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, pp 244-255, 1995.
- D. J. Hand and K. Yu. Idiot Bayes ? not so stupid after all? *International Statistical Review*, 69:385-398, 2001.
- C. N. Hsu, H. J. Huang and T. T Wong. Implications of the Dirichlet Assumption for Discretization of Continuous Variables in Naive Bayesian Classifiers. *Machine Learning*, 53(3):235-263, 2003.
- G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119-127, 1980.
- R. Kass and A. Raftery. Bayes factors. In *Journal of the American Statistical Association*, 90: 773-795, 1995.
- R. Kerber. Chimerge discretization of numeric attributes. *Proceedings of the 10th International Conference on Artificial Intelligence*, pp 123-128, 1991.
- S. Kullback. *Information Theory and Statistics*. New York: Wiley, (1959); republished by Dover, 1968.

- P. Langley, W. Iba and K. Thompson. An analysis of Bayesian classifiers. In *Proceedings of the 10th national conference on Artificial Intelligence*, AAAI Press, pp 223-228, 1992.
- Y. Lechevallier. Recherche d'une partition optimale sous contrainte d'ordre total. Technical report N°1247. INRIA, 1990.
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.
- J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81-106, 1986.
- J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
- G. Ritschard, D. A. Zighed and N. Nicoloyannis. Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé. *Mathématiques et Sciences Humaines*, n°154-155:81-98, 2001.
- G. Ritschard. Partition BIC optimale de l'espace des prédicteurs. *Revue des Nouvelles Technologies de l'Information*, 1:99-110, 2003.
- G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461-464, 1978.
- SPSS Inc. *AnswerTree 3.0 User's Guide*. Chicago: SPSS Inc, 2001.
- Y. Yang and G. Webb. On why discretization works for naïve-Bayes classifiers. *Proceedings of the 16th Australian Joint Conference on Artificial Intelligence (AI)*, 2003.
- D. A. Zighed and R. Rakotomalala. *Graphes d'induction*. Hermes Science Publications, pp 327-359, 2000.