A Robust Method for Partitioning the Values of Categorical Attributes

Marc Boullé *

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

Résumé. Dans le domaine de l'apprentissage supervisé, les méthodes de groupage des modalités d'un attribut symbolique permettent de construire un nouvel attribut synthétique conservant au maximum la valeur informationnelle de l'attribut initial et diminuant le nombre de modalités. Nous proposons ici une généralisation de l'algorithme de discrétisation Khiops[1] pour le problème du groupage des modalités. L'algorithme proposé permet de contrôler a priori le risque de sur-apprentissage et d'améliorer significativement la robustesse des groupages produits. Cette caractéristique de robustesse a été obtenue en étudiant la statistique des variations du critère du Khi2 lors de regroupements de lignes d'un tableau de contingence et en modélisant le comportement statistique de l'algorithme Khiops. Des expérimentations intensives ont permis de valider cette approche et ont montré que la méthode de groupage Khiops aboutit à des groupages performants, à la fois en terme de qualité prédictive et de faible nombre de groupes.

1. Introduction

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 datasets, there are many cases where the grouping of values of categorical attributes is a mandatory preprocessing step. 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 et Rakotomalala, 2000]. Neural nets are based on numerical 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, linear regression 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 [Cestnik *et al.*, 1987] and evaluated by [Berckman, 1995] is a

[1] French patents N° 01 07006 and N° 02 16733

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 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 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 probably be extended to find the optimal partition of the categories into K groups in the case of two class values, with the use 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 confidence level of 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 such as the Cramer's V or the Tschuprow's T [Ritschard *et al.*, 2001] in order to compare two different-size partitions.

The Khiops grouping method is a straightforward generalization of the Khiops discretization method [Boullé, 2003a]. Instead of merging adjacent numerical values in order to build intervals, the grouping method merges categorical values into groups of values. In both cases, the search algorithm is a bottom-up greedy heuristic that optimizes the chi-square criterion applied to the whole set of intervals or groups. The stopping rule is based on the confidence level computed with chi-square statistics. The method automatically stops the merging process as soon as the confidence level, related to the test of independence between the partitioned attribute and the class attribute, does not decrease anymore.

The set of groups resulting from a grouping method provides an elementary univariate classifier, which predicts the distribution of the class values in each learned group. A grouping method can be considered as an inductive algorithm, therefore subject to overfitting. We apply a methodology similar to that developed for the Khiops discretization method in order to bring a true control of overfitting. The principle is to analyze the behavior of the algorithm during the grouping of an explanatory attribute independent from the class attribute. We study the statistics of the variations of the chi-square values during the merge of categories and propose to model the maximum of these variations in a complete grouping process. The algorithm is then modified in order to force any merge whose variation of chi-square value is below the maximum variation predicted by our statistical modeling. This change in the algorithm yields the interesting probabilistic guarantee that any independent attribute will be grouped within a single terminal group and that any attribute whose

grouping consists of at least two groups truly contains predictive information upon the class attribute. This is experimentally confirmed.

The remainder of the document is organized as follows. Section 2 briefly introduces the initial Khiops grouping algorithm. Section 3 presents the statistical modeling of the algorithm and its fine-tuning to prevent overfitting. Section 4 proceeds with an extensive experimental evaluation.

2. The Khiops Grouping Method

In this section, we recall the principles of the chi-square test and present the Khiops grouping algorithm, whose detailed description and analysis can be found in [Boullé, 2003b].

2.1 The Chi-square Test: Principles and Notations

Let us consider an explanatory attribute and a class attribute and determine whether they are independent. First, all instances are summarized in a contingency table, where the instances are counted for each value pair of explanatory and class attributes. The chi-square value is computed from the contingency table, based on table 1 notations.

n_{ij}: Observed frequency for ith explanatory value and jth class value
$n_{i.}$: Total observed frequency for ith explanatory value
$n_{.j}$: Total observed frequency for jth class value
N: Total observed frequency
I: Number of explanatory attribute values
J: Number of class values

	A	B	C	Total
a	n_{11}	n_{12}	n_{13}	$n_{1.}$
b	n_{21}	n_{22}	n_{23}	$n_{2.}$
c	n_{31}	n_{32}	n_{33}	$n_{3.}$
d	n_{41}	n_{42}	n_{43}	$n_{4.}$
e	n_{51}	n_{52}	n_{53}	$n_{5.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

TAB 1 – *Contingency table used to compute the chi-square value.*

Let $e_{ij} = n_{i.}.n_{.j} / N$, stand for the expected frequency for cell (i, j) if the explanatory and class attributes are independent. The chi-square value is a measure on the whole contingency table of the difference between observed frequencies and expected frequencies. It can be interpreted as a distance to the hypothesis of independence between attributes.

$$Chi2 = \sum_i \sum_j \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \quad . \tag{1}$$

Within the null hypothesis of independence, the chi-square value is subject to chi-square statistics with (I-1).(J-1) degrees of freedom. This is the basis for a statistical test which allows to reject the hypothesis of independence; the higher the chi-square value is, the smaller the confidence level is.

2.2 Initial Algorithm

The chi-square value depends on the local observed frequencies in each individual row and on the global observed frequencies in the whole contingency table. This is a good

candidate criterion for a grouping method. The chi-square statistics is parameterized by the number of explanatory values (related to the degrees of freedom). In order to compare two groupings with different group numbers, we use the confidence level instead of the chi-square value.

The principle of the Khiops algorithm is to minimize the confidence level between the grouped explanatory attribute and the class attribute by the means of chi-square statistics. The chi-square value is not reliable to test the hypothesis of independence if the expected frequency in any cell of the contingency table falls below some minimum value. The algorithm copes with this constraint in a preprocessing step: any initial category that does not fulfill the minimum frequency constraint is unconditionally merged into a special group.

The Khiops method is based on a greedy bottom-up algorithm. It starts with initial categories and then searches for the best merge between categories. The algorithm is reiterated until no further merge can decrease the confidence level. The computational complexity of the algorithm can be reduced to $O(N.\log(N) + I^2.\log(I))$ with some optimizations [Boullé, 2003b].

There are two main differences between the initial Khiops algorithm and the similar CHAID algorithm [Kass, 1980]. First, the chi-square criterion is applied globally to the *whole partition* in the case of the Khiops algorithm, whereas it is applied locally to *two adjacent groups* in the case of the CHAID algorithm. Second, the Khiops algorithm stops the merging process when *the confidence level increases* after the best candidate merge, whereas the CHAID algorithm stops when *the confidence level is beyond a threshold* set by the user.

3. Statistical Analysis of the Algorithm

The Khiops algorithm chooses the best merge among all possible merges of categories and iterates this process until the stopping rule is met. When the explanatory attribute and the class attribute are independent, the resulting set of groups should be composed of a single group, meaning that there is no predictive information in the explanatory attribute. In the following, we study the statistical behavior of the initial Khiops algorithm.

In the case of two independent attributes, the chi-square value is subject to chi-square statistics, with known expectation and variance. We study the DeltaChi2 law (variation of the chi-square value after the merge of two categories) in the case of two independent attributes. During a grouping process, a large number of merges are evaluated, and at each step, the Khiops algorithm chooses the merge that maximizes the chi-square value; i.e. the merge that minimizes the DeltaChi2 value since the chi-square value before the merge is fixed. The stopping rule is met when the best DeltaChi2 value is too large. However, in the case of two independent attributes, the merging process should continue until the grouping algorithm reaches a single terminal group. The largest DeltaChi2 value encountered during the algorithm merging decision steps must then be accepted. We will try to estimate this MaxDeltaChi2 value in the case of two independent attributes and modify the algorithm in order to force the merges as long as this bound is not reached.

3.1 Statistics of the MaxDeltaChi2 Values of the Khiops Algorithm

Let us focus on two rows r and r' of the contingency table, with frequencies n and n', and row probabilities of the class values $p_1, p_2, \ldots p_J$ and $p'_1, p'_2, \ldots p'_J$. Let $P_1, P_2, \ldots P_J$ be the

probabilities of the class values on the whole contingency table. The chi-square value can only decrease when the two rows are merged. Let us define the DeltaChi2 value as the variation of the chi-square value during a merge.

$$DeltaChi2 = \frac{nn'}{n+n'} \sum_{j=1}^{J} \frac{\left(p_j - p'_j\right)^2}{P_j} \quad . \tag{2}$$

We proved that in the case of an explanatory attribute independent from a class attribute with *J* class values, the DeltaChi2 value resulting from the merge of two rows with the same frequencies is asymptotically distributed as the chi-square statistics with *J*-1 degrees of freedom [Boullé, 2003b].

The MaxDeltaChi2 value is equal to the maximum of the DeltaChi2 values encountered during the complete grouping process downward a single terminal group, when the grouped attribute is independent of the class attribute. In the case of a discretization process, where the merges are constrained to be adjacent in the contingency table, we proposed in [Boullé, 2003a] an analytic formula to approximate the statistics of the MaxDeltaChi2. In the case of a grouping process, we were not able to approximate the statistics of the MaxDeltaChi2 analytically. However, we showed in [Boullé, 2003b] that the statistics of the MaxDeltaChi2 depends only on two parameters: the number of initial categories *I* and the number of class values *J*. More precisely, the following propositions are conjectures that have been checked through extensive experiments on synthetic data:

- the statistics of the MaxDeltaChi2 is independent of the sample size,
- the statistics of the MaxDeltaChi2 is independent of the distribution of the categories,
- the statistics of the MaxDeltaChi2 is independent of the distribution of the class.

For example, the first conjecture was evaluated in the case of random datasets with 50 equidistributed initial categories and 2 equidistributed classes. We collected the MaxDeltaChi2 values resulting from a complete grouping process, for 1000 randomly generated datasets. This experiment was repeated for a large number of sample sizes ranging from 1000 to 200000 instances, and showed that the repartition functions of the MaxDeltaChi2 values are independent of the sample size. The same kind of experiments was performed to check the other conjectures.

We also proved the following propositions in the cases where there are only two categories or two classes.

Proposition 1. In the case of two categories and *J* classes, the statistics of the MaxDeltaChi2 value is the chi-square statistics with (*J*-1) degrees of freedom.

Proposition 2. In the case of *I* equidistributed categories and two equidistributed classes, the mean of the MaxDeltaChi2 value is asymptotically equal to $2I/\pi$.

In the general case, the statistics of the MaxDeltaChi2 value could not be modeled with a mathematical expression, like that of the Khiops discretization method. We choose to compute experimentally the mean and standard deviation of the MaxDeltaChi2, for a large number of pairs of parameters (*I*, *J*). The analysis of the results reveals a linear behavior with

respect to both parameters *I* and *J*, which is consistent with propositions 1 and 2. This observation allows to use a value table to approximate the mean and standard deviation of the MaxDeltaChi2 values and to rely on a linear interpolation between pre-computed values. Finally, we make a last assumption, confirmed by experimental evaluation: the repartition function of the MaxDeltaChi2 values can be approximated by a normal law with the same mean and standard deviation. Full details of the simulation are given in [Boullé, 2003b].

To conclude, the MaxDeltaChi2 value used by the Khiops grouping algorithm is calculated owing to a linear interpolation of the mean and standard deviation found in a pre-computed value table for given numbers of categories and of class values. Using the inverse normal law, the MaxDeltaChi2 value is determined so that it will be greater than the observed DeltaChi2 values with probability p (p=0.95 for instance).

3.2 The Robust Khiops Grouping Algorithm

Algorithm Robust Khiops
1. Initialization
   - 1.1 Sort the explanatory attribute values
   - 1.2 Create an elementary group for each value
   - 1.3 Create a special group to handle all initial categories that do not fulfill the minimum frequency constraint; if necessary, merge this special group with the least frequent remaining category
   - 1.4 Compute the MaxDeltaChi2 value related to the number of initial groups and of class values
2. Optimization of the grouping: repeat the following steps
   - 2.1 Evaluate all possible merges between pairs of groups
   - 2.2 Search for the best merge
   - 2.3 Merge and continue as long as one of the following conditions is relevant
     - The confidence level of the grouping decreases after the merge
     - The DeltaChi2 value of the best merge is below the MaxDeltaChi2 value

In the case of two independent attributes, the grouping should result in a single terminal group. For a given probability p, the statistical modeling of the Khiops algorithms provides a theoretical value MaxDeltaChi2(p) that will be greater than all the DeltaChi2 values of the merges completed during the grouping process, with probability p. The initial Khiops grouping algorithm is then modified in order to force all the merges whose DeltaChi2 value is smaller than MaxDeltaChi2(p). This ensures the expected behavior of the algorithm with probability p. In the case of two attributes with unknown dependency relationship, this enhancement of the algorithm guarantees that when the grouped attribute consists of at least two groups, the explanatory attribute truly holds information concerning the class attribute with probability higher than p. We suggest to set p=0.95, in order to ensure reliable grouping results.

The impact on the initial Khiops algorithm is restricted to the evaluation of the stopping rule and retains the supra-linear computational complexity of the algorithm.

4. Experiments

Dataset	Continuous Attributes	Nominal 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

TAB 2 – *Datasets.*

In our experimental study, we compare the Khiops grouping method with other supervised grouping algorithms on two criterions: predictive performance and number of groups. 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 [Zighed et Rakotomalala, 2000], where each grouping method is considered as an elementary inductive method which predicts the distribution of the class values in each learned groups.

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 to predictive quality of the groupings, we use the Kullback-Leibler divergence [Kullback, 1968] applied to compare the distribution of the class values estimated from the learning set (based on the learned groups) with the distribution of the class values observed on the test set (based on the initial values: the same for all the tested methods). For a given category, let p_j be the probability of the j^{th} class value estimated on the learning set (with the use of the group containing the category), and q_j be the probability of the j^{th} class value observed on the test set (using only the category). The Kullback-Leibler divergence between the estimated distribution and the observed distribution is:

$$D(p \,||\, q) = \sum_{j=1}^{J} p_j \log \frac{p_j}{q_j} \ . \qquad (3)$$

The global evaluation of the predictive quality is computed as the mean of the Kullback-Leibler divergence on the test set. In order to smooth the empirical distributions and to deal with zero probabilities, we use the Laplace's estimator. For other approaches for defining goodness-of-fit measures, see for example [Ritschard et Zighed, 2003].

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. In the experiments, we collect both the predictive quality results using the Kullback-Leibler divergence and the number of groups.

We gathered 12 datasets from U.C. Irvine repository [Blake et Merz, 1998], each dataset 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. Table 2 describes the datasets; the last column corresponds to the accuracy of the majority class.

The grouping methods studied in the comparison are:

- Khiops: the method described in this paper,
- Initial Khiops: the initial version of the method, described in section 2,
- CHAID: the grouping method used in the CHAID method [Kass, 1980],
- Tschuprow: the grouping method described for example in [Ritschard *et al.*, 2001],
- Gain Ratio: the grouping method used in the C4.5 method [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 Gain Ratio method is the only method based on entropy; the other methods use chi-square based criterions. The initial Khiops method applies the chi-square criterion on the whole contingency table and evaluates the partition with the related confidence level. The robust Khiops method enhances the initial Khiops algorithm by providing guarantees against overfitting. The Tschuprow method is also based on a global evaluation of the contingency table, but it uses the Tschuprow's T normalization of the chi-square value instead of the confidence level to evaluate the partitions. The CHAID method applies the chi-square criterion locally to two rows of the contingency table. For the CHAID method, the significance level is set to 0.95 for chi-square threshold, and the Bonferroni correction is not applied. We have re-implemented these alternative grouping approaches in order to eliminate any variance resulting from different cross-validation splits. The groupings are performed on the 230 attributes of the datasets, using a stratified tenfold cross-validation. In order to determine whether the performances are significantly different between the Khiops method and the alternative methods, the t-statistics of the difference of the results is computed. Under the null hypothesis, this value has a Student’s distribution with 9 degrees of freedom. The confidence level is set to 5% and a two-tailed test is performed to reject the null hypothesis.

4.1 Quality of the Groupings

The whole result tables are too large to be printed in this paper. The predictive quality results are summarized in table 3, which reports for each dataset the mean of the Kullback-Leibler divergences and the number of significant Khiops wins (+) and losses (-) for each method comparison. The results have been normalized using the Kullback-Leibler divergence evaluated when no grouping is done. The means are geometric means in order to focus on the ratios of performances between the tested methods.

The results show significant differences between the methods which allow to rank the tested methods. In a first cluster of method, the Khiops grouping method obtains the best results, followed by the initial Khiops grouping method and then by the CHAID method. The Khiops method gets significantly better results than the CHAID method for 24% of the grouped attributes, and significantly worse results for 7% of the attributes. In a second cluster of methods, the Tchuprow and Gain Ratio methods are clearly outperformed by the leading three methods. For example, the Khiops method surpasses the Gain Ratio method for 35% of the attributes, and is beaten for only 3.5% of the attributes.

Dataset	Khiops	Ini. Khiops			CHAID			Tschuprow			Gain Ratio		
			+	-		+	-		+	-		+	-
Adult	1.05	1.13	3	2	1.07	4	4	3.76	10	0	4.16	10	0
Australian	1.04	1.06	0	0	1.10	2	0	1.10	1	0	1.24	3	0
Breast	1.24	1.24	1	0	1.36	4	0	1.45	2	0	1.66	5	0
Crx	1.06	1.07	0	1	1.08	0	1	1.10	1	0	1.23	3	0
Heart	0.98	1.02	1	0	1.02	0	0	1.03	2	0	1.07	3	0
HorseColic	1.02	1.01	1	4	1.07	3	2	1.08	3	0	1.04	3	2
Ionosphere	1.07	1.03	1	3	1.13	7	1	1.06	2	2	1.08	3	4
Mushroom	1.10	1.24	4	2	1.21	6	2	2.29	11	1	2.60	11	1
TicTacToe	0.97	0.97	0	0	0.91	0	1	0.95	0	0	0.95	0	0
Vehicle	1.10	1.10	5	1	1.11	4	4	1.12	2	2	1.30	9	0
Waveform	0.92	0.99	13	0	1.01	19	0	1.48	30	0	1.47	30	0
Wine	1.23	1.20	0	1	1.37	6	1	1.24	1	1	1.23	0	1
Synthesis	1.04	1.07	29	14	1.10	55	16	1.35	65	6	1.42	80	8

TAB 3 – *Means of the predictive quality of the groupings, number of significant wins (+) and losses (-) per dataset for the Khiops method when compared to the alternative methods.*

To summarize, the predictive quality criterion suggests the following ranking of the tested methods: Khiops, Initial Khiops, CHAID, Tschuprow, Gain Ratio.

4.2 Size of the Grouping

The group number results are summarized in table 4.

The differences are very significant between the tested methods. The Tschuprow and Gain Ratio methods produce the smallest size groupings on average, at the expense of a low predictive quality. Among the high quality grouping methods, the Khiops method is a clear winner for the group number criterion, followed by the initial Khiops method and the CHAID method. The groupings produced by the Khiops method are *always* smaller than these produced by the CHAID method, and the differences are significant for 60% of the attributes. Although the Tschuprow and Gain Ratio methods obtain smaller groupings on average, the results are contrasted among the datasets. For almost one fourth of the attributes, the Khiops method gets significantly smaller groupings than the Gain Ratio method.

It is interesting to analyze more deeply the results of the waveform dataset, where about half of the attributes are noise attributes. An inspection of the groupings reveals that the

robust Khiops grouping method is the only method that correctly identifies the noise attributes with groupings reduced to only one group.

Dataset	Khiops	Ini. Khiops			CHAID			Tschuprow			Gain Ratio		
			+	-		+	-		+	-		+	-
Adult	3.67	3.99	5	0	4.83	11	0	2.05	2	10	2.33	2	10
Australian	1.91	2.19	6	1	2.19	4	0	2.19	3	1	2.36	7	1
Breast	2.60	2.83	3	0	4.16	9	0	1.98	1	7	1.98	1	7
Crx	1.93	2.16	5	1	2.18	3	0	2.15	3	1	2.42	8	2
Heart	1.91	2.27	5	0	2.14	4	0	2.11	3	1	2.08	3	1
HorseColic	1.87	2.20	11	0	2.24	10	0	2.03	7	4	2.03	8	4
Ionosphere	2.47	2.94	17	1	3.18	25	0	2.09	0	15	2.05	0	17
Mushroom	3.06	3.11	3	3	3.57	10	0	2.00	0	13	2.19	1	13
TicTacToe	2.03	2.03	0	0	2.11	1	0	2.00	0	0	2.00	0	0
Vehicle	3.50	3.90	7	0	4.84	17	0	2.58	0	11	2.85	3	11
Waveform	2.67	3.56	30	0	3.76	35	0	2.73	21	19	3.18	21	18
Wine	2.60	2.95	5	0	3.56	11	0	2.10	0	6	2.05	1	7
Synthesis	2.54	2.95	97	6	3.28	140	0	2.22	40	88	2.38	55	91

TAB 4 – *Means of the size of the groupings, number of significant wins (+) and losses (-) per dataset for the Khiops method when compared to the alternative methods.*

To summarize, the group number criterion suggests the following ranking of the tested methods: Tschuprow, Gain Ratio, Khiops, Initial Khiops, CHAID.

4.3 Bi-criteria Analysis of the Results

In order to better understand the relations between predictive quality and size of the groupings, we draw in figure 1 the global means of the results on a two-criteria plan with the group number on the x-coordinate and the predictive quality on the y-coordinate. For comparison purposes, we also report the results obtained by three alternative simple grouping methods:

- Mode: unsupervised bipartition of the categories with one group containing the mode, i.e. the most frequent category,
- Chi Single Value: bipartition of the categories with one category against all the others, selected using the chi-square criterion (one final merge is still possible),
- Exhaustive CHAID: bipartition of the categories obtained with the CHAID algorithm by forcing the merges until the partition contains at most two groups.

The three bipartition grouping methods are ranked as expected for the predictive quality criterion. The Tschuprow and Gain Ratio methods that are allowed to build partition with more than two groups do not obtain better results on predictive quality than the Exhaustive CHAID method. The cluster of the efficient methods (Khiops, Initial Khiops and CHAID) clearly takes benefit of multi-group partitions. Among these leading methods, the Khiops method dominates the others methods on both criteria. Lastly, considering the computational complexity of the algorithms, that of the optimized Khiops algorithm is $O(N.\log(N) +$

I^2.log(I)), whereas that of the other methods is O(N.log(N) + I^3). However, the difference in runtime is minor in many cases, when the number of categorical values I is very small.

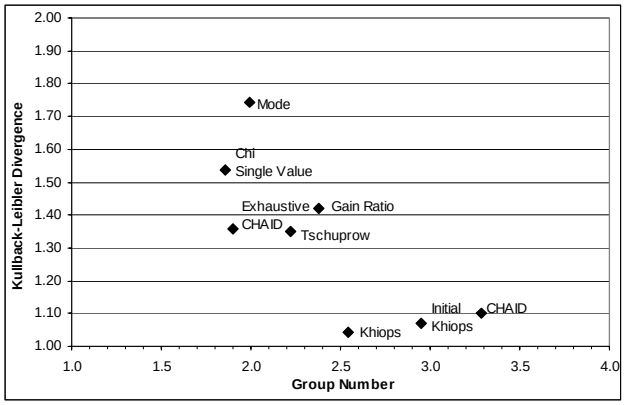

FIG. 1 - *Bi-criteria evaluation of the grouping methods for the group number and the predictive quality criteria*

5. Conclusion

The principle of the Khiops grouping method is to minimize the confidence level related to the test of independence between the grouped attribute and the class attribute. During the bottom-up process of the algorithm, numerous merges between categories are performed that produce variations of the chi-square value of the contingency table. Owing to a statistical modeling of these variations when the explanatory attribute is independent of the class attribute, we enhanced the initial Khiops grouping algorithm in order to guarantee that the groupings of independent attributes are reduced to a single group. This attested resistance to overfitting is an interesting alternative to the classical cross-validation approach.

Extensive comparative experiments show that the Khiops method outperforms the other tested grouping methods. It allows to drastically reduce the number of values of categorical attributes in the preprocessing step of data mining, while keeping most of their monothetic predictive performance.

Références

[Berckman, 1995] N.C. Berckman. Value grouping for binary decision trees. Technical Report. Computer Science Department – University of Massachusetts, 1995.

[Blake et Merz, 1998] C.L. Blake et C.J. Merz. UCI Repository of machine learning databases Web URL http://www.ics.uci.edu/~mlearn/MLRepository.html, Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[Boullé, 2003a] M. Boullé. Khiops: a Discretization Method of Continuous Attributes with Guaranteed Resistance to Noise. *Proceedings of the Third International Conference on Machine Learning and Data Mining in Pattern Recognition*, 50-64, 2003.

[Boullé, 2003b] M. Boullé. Groupage robuste des valeurs d'un attribut symbolique par la méthode Khiops. Note technique NT/FTR&D/8028, France Telecom R&D, 2003.
[Breiman et al., 1984] L. Breiman, J.H. Friedman, R.A. Olshen et C.J. Stone. Classification and Regression Trees. California: Wadsworth International, 1984.
[Cestnik et al., 1987] B. Cestnik, I. Kononenko et I. Bratko. ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In Bratko & Lavrac (Eds.), Progress in Machine Learning, Wilmslow, UK: Sigma Press, 1987.
[Chou, 1991] P.A. Chou. Optimal Partitioning for Classification and Regression Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):340-354, 1991.
[Fulton et al., 1995] T. Fulton, S. Kasif et S. Salzberg. Efficient algorithms for finding multi-way splits for decision trees. *Proceeding of the Thirteenth International Joint Conference on Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, 244-255, 1995.
[Kass, 1980] G.V. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119-127, 1980.
[Kerber, 1991] R. Kerber. Chimerge discretization of numeric attributes. *Proceedings of the 10th International Conference on Artificial Intelligence*, 123-128, 1991.
[Kullback, 1968] S. Kullback. Information Theory and Statistics. New York: Wiley, (1959); republished by Dover, 1968.
[Lechevallier , 1990] Y. Lechevallier. Recherche d'une partition optimale sous contrainte d'ordre total. Technical report N°1247, INRIA, 1990.
[Pyle, 1999] D. Pyle. Data Preparation for Data Mining, Morgan Kaufmann, 1999.
[Quinlan, 1986] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81-106, 1986.
[Quinlan, 1993] J.R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
[Ritschard et al., 2001] G. Ritschard, D.A. Zighed et N. Nicoloyannis. Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé. *Math. & Sci. Hum.*, n°154-155:81-98, 2001.
[Ritschard et Zighed, 2003] G. Ritschard et D.A. Zighed. Modélisation de tables de contingence par arbre d'induction. *Extraction et Gestion des Connaissances*, 381-392, 2003.
[Zighed et Rakotomalala, 2000] D.A. Zighed et R. Rakotomalala. Graphes d'induction. Hermes Science Publications, 327-359, 2000.

Summary

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 Khiops, based on a generalization of the Khiops discretization algorithm. This grouping method provides guarantees against overfitting and thus leads to robust groupings. This property derives from a statistical modeling of the Khiops method which allows to fine-tune the algorithm. Extensive experiments demonstrate the validity of this approach and show that the Khiops grouping method builds high quality groupings, both in terms of predictive quality and of small number of groups.