Khiops: a Discretization Method of Continuous Attributes with Guaranteed Resistance to Noise

Marc Boullé

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

Abstract. In supervised machine learning, some algorithms are restricted to discrete data and need to discretize continuous attributes. The Khiops* discretization method, based on chi-square statistics, optimizes the chi-square criterion in a global manner on the whole discretization domain. In this paper, we propose a major evolution of the Khiops algorithm, that provides guarantees against overfitting and thus significantly improve the robustness of the discretizations. This enhancement is based on a statistical modeling of the Khiops algorithm, derived from the study of the variations of the chi-square value during the discretization process. This modeling, experimentally checked, allows to modify the algorithm and to bring a true control of overfitting. Extensive experiments demonstrate the validity of the approach and show that the Khiops method builds high quality discretizations, both in terms of accuracy and of small interval number.

1 Introduction

Discretization of continuous attributes is a problem that has been studied extensively in the past [5,9,10,13]. Many induction algorithms rely on discrete attributes and need to discretize continuous attributes, i.e. to slice their domain into a finite number of intervals. For example, decision tree algorithms exploit a discretization method to handle continuous attributes. C4.5 [11] uses the information gain based on Shannon entropy. CART [4] applies the Gini criterion (a measure of the impurity of the intervals). CHAID [7] relies on a discretization method close to ChiMerge [8]. SIPINA takes advantage of the Fusinter criterion [12] based on measures of uncertainty that are sensitive to sample size. The Minimum Description Length Principle [6] is an original approach that attempts to minimize the total quantity of information both contained in the model and in the exceptions to the model.

The Khiops discretization method [2] is a bottom-up method based on the global optimization of chi-square. The Khiops method starts the discretization from the elementary single value intervals. It evaluates all merges between adjacent intervals and selects the best one according to the chi-square criterion applied to the whole set

* French patents N° 01 07006 and N° 02 16733

of intervals. The stopping rule is based on the confidence level computed with chi-square statistics. The method automatically stops merging intervals as soon as the confidence level, related to the chi-square test of independence between the discretized attribute and the class attribute, does not decrease anymore. The Khiops method optimizes a global criterion which evaluates the entire partition of the domain into intervals and not a local criterion applied to two neighboring intervals as in the ChiSplit top down method or the ChiMerge bottom-up method.

The set of intervals resulting from a discretization provides an elementary univariate classifier, which predicts the local majority class in each learned interval. A discretization method can be considered as an inductive algorithm, therefore subject to overfitting. This overfitting problem has not yet been deeply analyzed in the field of discretization. The initial Khiops discretization uses a heuristic control of overfitting by constraining the frequency of the intervals to be greater than the square root of the sample size. In this paper, we introduce a significant improvement of the Khiops algorithm which brings a true control of overfitting. The principle is to analyze the behavior of the algorithm during the discretization 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 intervals and propose a modeling of the maximum of these variations in a complete discretization 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 discretized within a single terminal attribute and that any attribute whose discretization consists of at least two intervals truly contains predictive information upon the class attribute.

The remainder of the document is organized as follows. Section 2 briefly introduces the initial Khiops algorithm. Section 3 presents the statistical modeling of the algorithm and its evolution. Section 4 proceeds with an extensive experimental evaluation.

2 The Initial Khiops Discretization Method

In this section, we recall the principles of the chi-square test and present the Khiops algorithm, whose detailed description and analysis can be found in [3].

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.

Table 1. Contingency table used to compute the chi-square value

n_{ij}: Observed frequency for i[th] explanatory value and j[th] class value
$n_{i.}$: Total observed frequency for i[th] explanatory value
$n_{.j}$: Total observed frequency for j[th] 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

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 $(I\text{-}1).(J\text{-}1)$ 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, the smaller the confidence level.

2.2 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 discretization method. The chi-square statistics is parameterized by the number of explanatory values (related to the degrees of freedom). In order to compare two discretizations with different interval numbers, we use the confidence level instead of the chi-square value.

The principle of Khiops algorithm is to minimize the confidence level between the discretized 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.

The Khiops method is based on a greedy bottom-up algorithm. It starts with initial single value intervals and then searches for the best merge between adjacent intervals. Two different types of merges are encountered. First, merges with at least one interval that does not meet the constraint and second, merges with both intervals fulfilling the constraint. The best merge candidate (with the highest chi-square value) is chosen in priority among the first type of merges (in which case the merge is accepted unconditionally), and otherwise, if all minimum frequency constraints are respected, it is selected among the second type of merges (in which case the merge is accepted under the condition of improvement of the confidence level). The algorithm is

reiterated until both all minimum frequency constraints are respected and no further merge can decrease the confidence level.

The computational complexity of the algorithm can be reduced to O(N log(N)) with some optimizations [3].

2.3 Minimum Frequency per Interval

In order to be reliable, the chi-square test requires that every cell of the contingency table have an expected value of at least 5. This is equivalent to a minimum frequency constraint for each interval of the discretization. Furthermore, to prevent overfitting, the initial Khiops algorithm heuristically increases the minimum frequency per interval constraint up to the square root of the sample size. In this paper, we show how to replace this heuristic solution by a method with theoretical foundations to avoid overfitting.

3 Statistical Analysis of the Algorithm

The Khiops algorithm chooses the best merge among all possible merges of intervals and iterates this process until the stopping rule is met. When the explanatory attribute and the class attribute are independent, the resulting discretization should be composed of a single interval, 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 intervals) in the case of two independent attributes. During a discretization 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 discretization reaches a single terminal interval. 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 The DeltaChi2 Law

The expectation and the variance of chi-square statistics with k degrees of freedom and a sample of size N are:

$E(Chi2) = k$,

$$Var(Chi2) = 2k + \frac{1}{N}\left(\sum_{i=1}^{k}\frac{1}{q_i} - k^2 - 4k - 1\right) .$$

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 $p'_1, p'_2, \ldots p'_J$.

					Total
	…	…	…	…	…
	…	…	…	…	
row r	$p_1 n$	$p_2 n$	…	$p_J n$	n
row r'	$p'_1 n'$	$p'_2 n'$	…	$p'_J n'$	n'
	…	…	…	…	
	…	…	…	…	
Total	$P_1 N$	$P_2 N$	…	$P_J N$	N

Owing to the additivity of the chi-square criterion, the variation of the chi-square value is based on the row contribution of the two rows before and after the merge.

$$Chi2_{afterMerge} - Chi2_{beforeMerge} = Chi2(r \cup r') - Chi2(r) - Chi2(r') .$$

$$Chi2_{afterMerge} - Chi2_{beforeMerge} = -\frac{nn'}{n+n'}\sum_{j=1}^{J}\frac{(p_j - p'_j)^2}{P_j} .$$

This variation of the chi-square value is always negative and is equal to zero only when the two rows hold exactly the same proportions of class values. The chi-square value of a contingency table can only decrease when two rows are merged. In the following, we define the DeltaChi2 value with the absolute value of the variation of the chi-square value for ease of use.

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

We proved in [3] 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. Under these assumptions, we can derive the following properties of the DeltaChi2 statistics.

$$p(DeltaChi2_J \geq x) \sim p(Chi2_{J-1} \geq x) .$$

$$E(DeltaChi2_J) \sim J - 1 .$$

$$V(DeltaChi2_J) \sim 2(J-1) .$$

3.2 Statistics of the Merges of the Khiops Algorithm

During the complete discretization process toward a single terminal interval, the number of merges is equal to the sample size. A straightforward modeling of the Khiops algorithm is that all these merges are equi-distributed, independent and that

they follow the theoretical DeltaChi2 statistics. This is an approximate modeling, mainly for the following reasons:

- the merges are not independent,
- the DeltaChi2 statistics is valid only asymptotically and for intervals with the same frequency,
- the Khiops algorithms uses a minimum frequency constraint that induces a hierarchy among the possible merges,
- at each step, the completed merge is the best one among all possible merges.

In order to evaluate this statistical modeling of the Khiops algorithm, we proceed with an experimental study. This experiment consists in discretizing an explanatory continuous attribute independent of a class attribute whose two class values are equi-distributed. In order to draw their repartition function, all the DeltaChi2 values associated with the completed merges are collected until one terminal interval is built. This process is applied on samples of sizes 100, 1000 and 10000. The resulting empirical repartition functions of the DeltaChi2 values are displayed on figure 1 and compared with the theoretical DeltaChi2 repartition function.

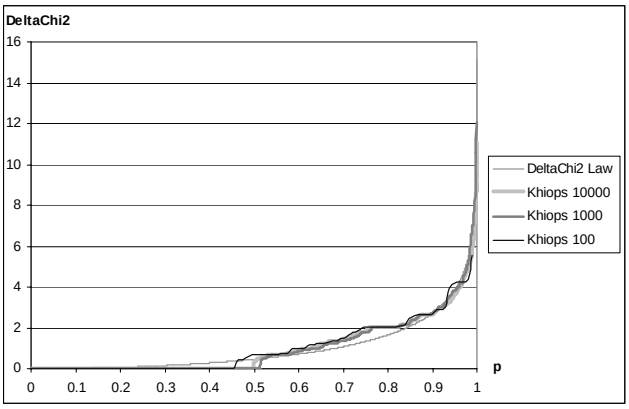

Fig. 1. Repartition functions of the DeltaChi2 values of the merges completed by the Khiops algorithm during the discretization of an explanatory attribute independent of the class attribute

The experiment shows that the DeltaChi2 empirical law is independent of the sample size and fits well the theoretical DeltaChi2 law, especially above the value p~0.85.

3.3 Statistics of the MaxDeltaChi2 Values of the Khiops Algorithm

The purpose is to settle a MaxDeltaChi2 threshold for the Khiops algorithm, so that in the case of two independent attributes, the algorithm converges toward a single terminal interval with a given probability p (p=0.95 for instance). All evaluated merges must be accepted as long as their DeltaChi2 value is below the MaxDeltaChi2 value. Based on the previous modeling where all the merges are independent, the

probability that all the merges are accepted is equal to the probability that one merge is accepted, to the power N.

The MaxDeltaChi2 value is given by:

$P(DeltaChi2_I \leq MaxDeltaChi2)^N \geq p$.

Using the theoretical DeltaChi2 law:

$P(Chi2_{J-1} \leq MaxDeltaKhi2) \geq p^{1/N}$.

$$MaxDeltaChi2 = InvChi2_{J-1}\left(prob \geq p^{1/N}\right) . \quad \textbf{(3)}$$

In order to validate this modeling of the MaxDeltaChi2 statistics, we proceed with a new experiment and collect the MaxDeltaChi2 value instead of all the DeltaChi2 values encountered during the algorithm. The experiment is applied on the same two independent attributes, on samples of sizes 100, 1000, 10000 and 100000. In order to gather many MaxDeltaChi2 values, it is repeated 1000 times for each sample size. The empirical MaxDeltaChi2 repartition functions are drawn on figure 2 and compared with the theoretical repartition functions derived from equation 3.

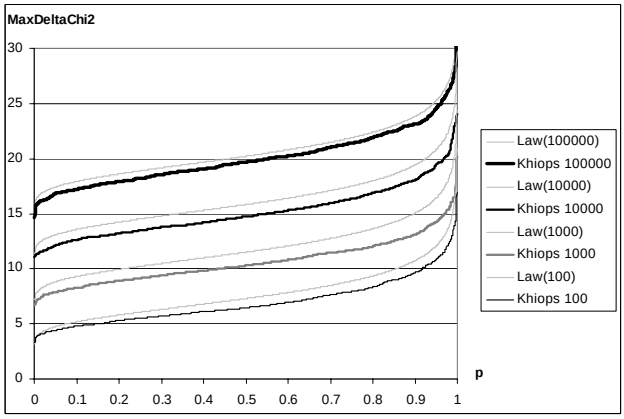

Fig. 2. Empirical and theoretical repartition function of the MaxDeltaChi2 values

The empirical and theoretical repartition functions have very similar shapes for each sample size. The theoretical values are upper bounds of the empirical values, with a moderate margin. We must keep in mind that these theoretical values result from an approximate statistical modeling of the Khiops algorithm. Their behavior as upper bounds is not proved but merely empirically observed.

3.4 The Robust Khiops Algorithm

The Khiops algorithm performs the merges of intervals as long as the confidence level of the chi-square test decreases. We keep the constraint of minimum frequency of 5 in each cell of the contingency table to ensure the reliability of the chi-square test, but

we replace the former heuristic minimum frequency constraint used to prevent overfitting by a new method based on the study of the MaxDeltaChi2 statistics.

In the case of two independent attributes, the discretization should result in a single terminal interval. 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 discretization, with probability p (probability higher than p according to the experimental study). The Khiops 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 discretized attribute consists of at least two intervals, 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 discretization results.

Algorithm Robust Khiops
1. Initialization
    1.1 Compute the MaxDeltaChi2 value with formula 3
    1.2 Sort the explanatory attribute values
    1.3 Create an elementary interval for each value
2. Optimization of the discretization: repeat the following steps
    2.1 Evaluate all possible merges between adjacent intervals
    2.2 Search for the best merge
    2.3 Merge and continue as long as one of the following conditions is relevant
        - At least one interval does not respect the minimum frequency constraint
        - The confidence level of the discretization decreases after the merge
        - The DeltaChi2 value of the best merge is below the MaxDeltaChi2 value

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

3.5 Post-optimization of the Discretizations

The Khiops method is a greedy bottom-up algorithm that allows identifying fine grain structures within efficient computation time. We propose a very simple post-processing of the discretizations in order to refine the boundaries of the intervals. For each pair of adjacent intervals, the post processing searches for the best boundary between the two intervals. This local optimization step is reiterated on all the pairs of intervals of the whole discretization, until no more improvement can be found. Experiments showed that this elementary post-optimization of the discretizations repeatedly brought slight improvements.

4 Experiments

In our experimental study, we compare the Khiops method with other supervised and unsupervised discretization algorithms. In order to evaluate the intrinsic performance of the discretization methods and eliminate the bias of the choice of a specific induction algorithm, we use a protocol similar as [13], where each discretization method is considered as an elementary inductive method, that predicts the local majority class in each learned interval. The discretizations are evaluated for two criteria: accuracy and interval number.

We gathered 15 datasets from U.C. Irvine repository [1], each dataset has at least one continuous attribute and at least a few tenths of instances for each class value. Table 2 describes the datasets; the last column corresponds to the accuracy of the majority class.

Table 2. Datasets

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
German	24	0	1000	2	70.00
Heart	10	3	270	2	55.56
Hepatitis	6	13	155	2	79.35
Hypothyroid	7	18	3163	2	95.23
Ionosphere	34	0	351	2	64.10
Iris	4	0	150	3	33.33
Pima	8	0	768	2	65.10
SickEuthyroid	7	18	3163	2	90.74
Vehicle	18	0	846	4	25.77
Waveform	21	0	5000	3	33.92
Wine	13	0	178	3	39.89

The discretization methods studied in the comparison are:

- Khiops: the method described in this paper,
- Initial Khiops: the previous version of the method, described in section 2,
- MDLPC: Minimum Description Length Principal Cut [6],
- ChiMerge: bottom-up method based on chi-square [8],
- ChiSplit: top-down method based on chi-square,
- Equal Width,
- Equal Frequency.

The MDLPC and initial Khiops methods have an automatic stopping rule and do not require any parameter setting. For the ChiMerge and ChiSplit methods, the significance level is set to 0.95 for chi-square threshold. For the Equal Width and Equal Frequency unsupervised discretization methods, the interval number is set to 10. We have re-implemented these alternative discretization approaches in order to

eliminate any variance resulting from different cross-validation splits. The discretizations are performed on the 181 single continuous 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 Accuracy of Discretizations

The whole result tables are too large to be printed in this paper. The accuracy results are summarized in table 3, which reports for each dataset the mean of the dataset attribute accuracies and the number of significant Khiops wins (+) and losses (-) of the elementary attribute classifiers for each method comparison. The results show that the supervised methods (except ChiMerge) perform clearly better than the unsupervised methods. The ChiMerge method is slightly better than the EqualWidth method, but not as good as the EqualFrequency method. The MDLPC method is clearly better than the EqualFrequency, ChiMerge and EqualWidth methods. The modified Khiops method outperforms the initial Khiops method. The Khiops and the ChiSplit methods obtain the best results of the experiment.

Table 3. Means of accuracies, number of significant wins and losses per dataset, for the elementary attribute classifiers

Dataset	Khiops	Init. Khiops			MDLPC			ChiMerge			ChiSplit			Eq. Width			Eq. Freq.		
			+	-		+	-		+	-		+	-		+	-		+	-
Adult	77.3	77.2	2	1	77.3	0	2	75.7	2	2	77.3	0	2	76.8	2	1	76.6	2	1
Australian	64.8	64.5	1	0	65.0	0	0	64.7	0	0	65.1	0	0	61.4	3	0	65.7	0	0
Breast	85.8	86.0	0	1	86.1	0	1	85.6	0	1	85.9	0	1	86.0	0	1	85.7	1	1
Crx	65.0	64.5	0	0	65.2	0	0	63.8	2	0	65.3	0	0	61.1	3	0	65.6	0	1
German	70.1	70.0	0	0	70.0	0	0	70.0	0	0	70.1	0	0	70.1	0	2	70.0	0	0
Heart	64.4	63.8	0	0	64.0	0	0	64.0	0	0	63.8	0	0	63.9	2	0	64.5	1	0
Hepatitis	79.6	79.4	0	0	79.3	0	0	77.8	3	0	79.3	0	0	79.8	0	0	79.9	0	0
Hypothyroid	96.1	96.0	0	1	96.1	0	1	96.0	3	0	96.1	1	0	95.4	3	1	95.2	3	1
Ionosphere	79.7	78.7	5	0	77.6	10	2	75.7	21	0	79.5	4	3	73.9	19	1	75.0	22	0
Iris	78.8	77.7	0	0	75.5	1	0	77.0	0	0	78.8	0	0	76.5	1	0	76.3	0	0
Pima	66.3	66.8	1	1	66.1	0	0	65.6	2	0	66.5	0	0	66.8	0	1	66.3	0	1
SickEuthyroid	91.3	91.4	0	0	91.3	0	0	91.3	1	0	91.3	0	0	90.7	2	0	91.0	1	0
Vehicle	41.5	40.9	3	1	40.5	4	0	41.4	2	1	42.1	0	3	40.8	3	0	40.3	3	0
Waveform	49.3	49.1	2	0	49.3	0	0	48.7	6	0	49.1	4	0	49.2	3	3	49.5	1	4
Wine	60.0	62.0	1	3	60.1	0	1	59.6	1	0	60.4	0	1	61.4	2	2	60.8	1	2
Synthesis	68.6	68.4	15	8	68.0	15	7	67.4	43	4	68.6	9	10	67.2	43	12	67.6	35	11

A close look at table 3 indicates a special behaviour of the ionosphere dataset, where the Khiops and ChiSplit methods largely dominate the other methods. An

inspection of the discretizations performed by the Khiops algorithm reveals unbalanced sets of intervals and non-monotonic distributions. The unsupervised methods cannot match the changes in the distributions since they cannot adjust the boundaries of the intervals. Furthermore, the discretizations present a frequent pattern consisting of an interesting interval nested between two regular intervals. This kind of pattern, easily detected by the Khiops bottom-up approach, is harder to discover for the MDLPC top-down algorithm since it requires two successive splits with the first one not very significant. The ChiSplit top-down method, which produces twice the interval number of the MDLPC method, manages to detect the interesting patterns at the expense of some unnecessary intervals. The ChiMerge method generates far too many intervals and over-learns the attributes.

All in all, the differences of accuracy may seem unimportant, but they are significant and must be compared to the average accuracy of the majority class classifier, which is 57.4%. Furthermore, the performances are averaged on a large variety of explanatory continuous attributes. It is interesting to analyze the differences of accuracy for the 181 attributes in more details. Figure 3 shows the repartition function of the differences of accuracy between the Khiops methods and the other discretization methods.

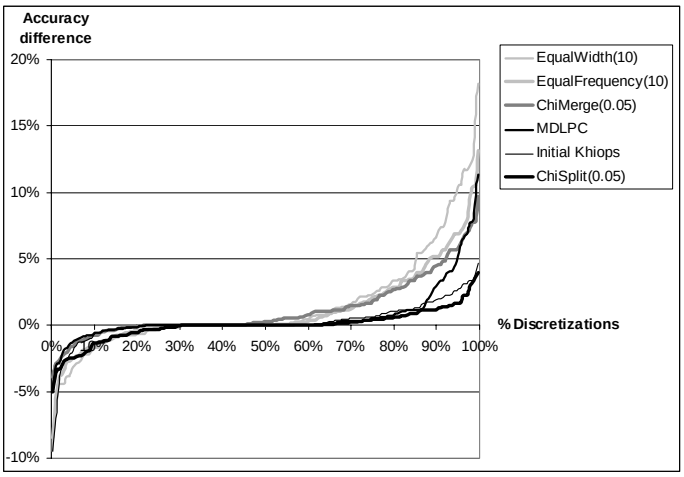

Fig. 3. Repartition function of the differences of accuracy between the Khiops method and the other discretization methods

On the left of the figure, the Khiops method is dominated by the other methods and, on the right, it outperforms the other algorithms. For about 40% of the attributes (between x-coordinates 20 and 60), all the discretization methods obtain equivalent results. Compared to the MDLPC method, the Khiops method is between 0 and 3% less accurate in about 10% of the discretizations, but is between 3 and 10% more accurate in about 10% of the discretizations. The average difference of 0.6% is thus significant and reflects potential large differences of accuracy on individual attributes.

The accuracy criterion suggests the following ranking of the tested methods:

1. Khiops, ChiSplit
2. Initial Khiops, MDLPC
3. EqualFrequency, ChiMerge
4. EqualWidth

4.2 Interval Number of Discretizations

The interval number results are summarized in table 4. The EqualWidth and EqualFrequency methods do not always reach the 10 required intervals, for reasons of lack of distinct explanatory values. The Khiops and MDLPC methods produce small size discretizations and are not significantly different for this criterion. The modified Khiops method generates almost half the interval number of the initial Khiops method. The ChiSplit method builds discretization with more than twice the interval number of the Khiops method. The ChiMerge method generates considerable interval numbers, especially for the larger samples.

Table 4. Means of interval numbers, number of significant wins and losses per dataset, for the elementary attribute classifiers

Dataset	Khiops	Init. Khiops			MDLPC			ChiMerge			ChiSplit			Eq. Width			Eq. Freq.		
			+	-		+	-		+	-		+	-		+	-		+	-
Adult	8.5	20.8	2	4	8.8	2	3	1264	0	7	28.2	0	6	9.4	2	4	6.6	4	2
Australian	2.1	5.3	0	6	2.0	0	0	16.1	0	6	5.2	0	6	8.1	0	6	8.8	0	6
Breast	2.6	3.7	0	5	2.9	0	5	11.6	0	9	4.9	0	9	9.2	0	10	5.9	0	10
Crx	2.1	5.3	0	6	2.1	0	0	15.8	0	6	5.1	0	6	8.2	0	6	8.7	0	6
German	1.3	2.6	0	20	1.2	2	0	2.4	0	12	2.0	0	12	3.8	0	23	3.4	0	21
Heart	1.7	3.1	0	6	1.7	0	0	5.0	0	5	2.5	0	5	5.9	0	8	6.1	0	8
Hepatitis	1.7	2.6	1	4	1.4	1	0	6.4	0	6	2.8	0	5	8.6	0	6	9.2	0	6
Hypothyroid	3.5	4.3	3	3	3.1	3	0	15.3	0	7	6.0	0	7	9.6	0	7	8.3	0	7
Ionosphere	4.3	5.1	3	21	3.9	11	6	30.0	0	32	8.0	0	30	9.4	0	32	8.9	0	32
Iris	2.8	3.3	0	2	2.8	1	0	3.7	0	3	3.6	0	3	9.7	0	4	9.5	0	4
Pima	2.3	4.5	0	8	2.1	2	1	13.2	0	8	5.0	0	8	9.5	0	8	9.3	0	8
SickEuthyroid	3.4	7.7	0	6	3.0	3	0	17.2	0	6	5.8	0	5	9.6	0	7	8.3	0	6
Vehicle	4.0	5.8	0	16	3.9	3	3	9.7	0	18	8.0	0	18	9.6	0	18	9.6	0	18
Waveform	4.5	9.4	1	19	4.9	1	9	49.0	0	21	13.6	0	21	10.0	0	21	10.0	0	21
Wine	2.6	3.5	0	11	2.8	1	3	6.7	0	13	4.8	0	12	9.8	0	13	10.0	0	13
Synthesis	3.3	5.6	10	137	3.2	30	30	66.1	0	159	7.2	0	153	8.5	2	173	8.0	4	168

The interval number criterion suggests the following ranking of the tested supervised methods:

1. Khiops, MDLPC
2. Initial Khiops
3. ChiSplit
4. ChiMerge

4.3 Multi-criteria Analysis of the Performances

The preceding results allow the ranking of the tested discretization methods on each criterion. It is interesting to use multi-criteria methodology to better understand the relations between accuracy and interval number. Let us recall some principles of multi-criteria analysis. A solution *dominates* (or is *non-inferior* to) another one if it is better for all criteria. A solution that cannot be dominated is *Pareto optimal*: any improvement on one of the criteria causes deterioration on another criterion. The *Pareto surface* (Pareto curve for two criteria) is the set of all the Pareto optimal solutions.

In order to study the importance of the parameters, we proceed with the previous experiments, using a wide range of parameters for each method. Figure 4 summarizes all the results on a two-criteria plan with the accuracy on the x-coordinate and the interval number on the y-coordinate.

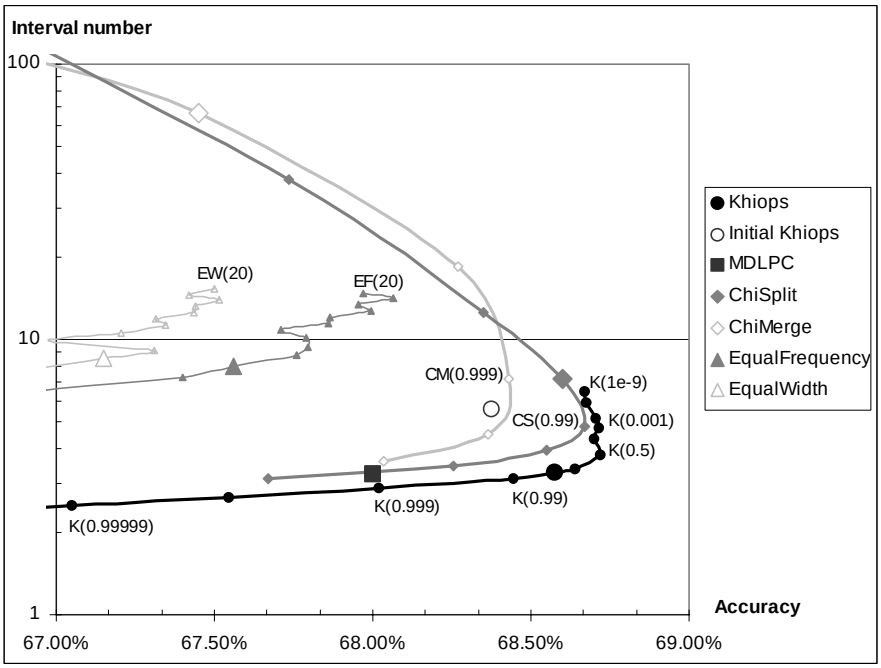

Fig. 4. Bi-criteria evaluation of the discretization methods for the accuracy and the interval number. The curves show the impact of the parameters on the performances for each method. The default parameters are located on the large size symbols

The unsupervised EqualWidth and EqualFrequency methods are largely dominated by the supervised methods. The ChiMerge method is the least performant among the supervised methods, especially for the interval number criterion. With its parameter set to 0.95, it produces 66 intervals on average. The MDLPC method builds very few intervals, but it is outperformed on accuracy by the other supervised methods,

principally the Khiops and ChiSplit methods, since their parameter can be tuned. The ChiSplit method exhibits high level performances on both criteria, but it is extremely sensitive to its parameter. Its top accuracy reaches that of the Khiops method, but it always needs significantly more intervals to obtain the same level of accuracy.

The Khiops method obtains the best results for both criteria and its curve corresponds to the Pareto curve for all the tested methods. For example, with Khiops parameter set to 0.95, the MDLPC methods constructs a similar interval number but is significantly outperformed on accuracy, whereas the ChiSplit method (with best parameter 0.99) achieves the same level of accuracy with a notably greater interval number. Compared to the initial Khiops method, the changes in the robust version of the algorithm bring notable enhancements on both criteria.

The ChiMerge and ChiSplit methods display similar curves on the two-criteria plan. With very strict parameters (probability almost equal to 1), they produce few intervals at the expense of a low accuracy. The interval number and the accuracy increase when the parameter is slightly relaxed, until a maximum is reached (with parameter 0.99 for ChiSplit and 0.999 for ChiMerge). Beyond this parameter threshold, the two methods are clearly subject to overfitting and display an increasing interval number associated with a deteriorating accuracy.

The Khiops method displays a steady behavior in the range of parameters between 0.95 and 0.5. With conservative parameters (probability almost equal to 1), it produces few intervals with poor accuracy. When the parameter moves to the "reasonable" range around 0.95, the accuracy quickly improves with a marginal increase of the interval number. After a maximum around parameter 0.5, the decrease of the parameter involves an increasing interval number, but surprisingly no decay in accuracy. An analysis of this behavior shows that the new intervals correspond to small statistical fluctuations whose cumulated effect on accuracy is not meaningful.

To conclude, the Khiops method demonstrates the best trade off between accuracy and interval number and has a stable behavior concerning its parameter. On the range of parameters between 0.95 and 0.5, the Khiops method dominates all the other tested discretization methods whatever the choice of the parameter.

5 Conclusion

The principle of the Khiops discretization method is to minimize the confidence level related to the test of independence between the discretized explanatory attribute and the class attribute. During the bottom-up process of the algorithm, numerous merges between intervals 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 Khiops algorithm in order to guarantee that the discretizations of independent attributes are reduced to a single interval. 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 discretization methods. A multi-criteria analysis of the results in terms of accuracy and interval number is very instructive and reveals an interesting behavior of

the Khiops algorithm, whose accuracy does not decrease even when the choice of its parameter might cause over-learning.

The Khiops method is an original approach that incorporates struggle against overfitting in its algorithm and exhibits both a high accuracy and small size discretizations.

Acknowledgement

I wish to thank Fabrice Clérot and Jean-Emmanuel Viallet for many insightful discussions and careful proof reading of the manuscript.

References

1. Blake, C.L., Merz, C.J.: 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)
2. Boullé, M.: Khiops: une méthode statistique de discrétisation. Extraction des connaissances et apprentissage, Vol 1-n°4. Hermes Science Publications (2001) 107-118
3. Boullé, M.: Amélioration de la robustesse de la méthode Khiops par contrôle de son comportement statistique. Note technique NT/FTR&D/7864. France Telecom R&D (2002)
4. Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J.: Classification and Regression Trees. California: Wadsworth International (1984)
5. Dougherty, J., Kohavi, R and Sahami, M.: Supervised and Unsupervised Discretization of Continuous Features. Proceedings of the Twelf International Conference on Machine Learning, Los Altos, CA: Morgan Kaufmann, (1995) 194-202
6. Fayyad, U., Irani, K.: On the handling of continuous-valued attributes in decision tree generation. Machine Learning, 8 (1992) 87-102
7. Kass, G.V.: An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29(2) (1980) 119-127
8. Kerber, R.: Chimerge discretization of numeric attributes. Proceedings of the 10th International Conference on Artificial Intelligence (1991) 123-128
9. Liu, H., Hussain, F., Tan, C.L. and Dash, M. Discretization: An Enabling Technique. Data Mining and Knowledge Discovery 6 (4) (2002) 393-423
10. Perner, P., Trautzsch, S.: Multi-interval Discretization Methods for Decision Tree Learning. SSPR/SPR (1998) 475-482
11. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
12. Zighed, D.A., Rabaseda, S. & Rakotomalala, R.: Fusinter: a method for discretization of continuous attributes for supervised learning. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(33) (1998) 307-326.
13. Zighed, D.A., Rakotomalala, R.: Graphes d'induction. Hermes Science Publications (2000) 327-359