Bias Correction in Classification Tree Construction

Alin Dobra DOBRA@CS.CORNELL.EDU
Johannes Gehrke JOHANNES@CS.CORNELL.EDU
Computer Science Department, Cornell University, Ithaca, NY 14853 USA

Abstract

We address the problem of bias in split variable selection in classification tree construction. A split criterion is unbiased if the selection of a split variable X is based only on the strength of the dependency between X and the class label, regardless of other characteristics (such as the size of the domain of X); otherwise the split criterion is biased. Our work makes the following four contributions: (1) We give a definition that allows us to quantify the extent of the bias of a split criterion, (2) we show that the p-value of any split criterion is a nearly unbiased criterion, (3) we give theoretical and experimental evidence that the correction is successful, and (4) we demonstrate the power of our method by correcting the bias of the `gini` gain.

1. Introduction

Split variable selection is one of the main components of classification tree construction. The quality of the split selection criterion has a major impact on the quality (generalization, interpretability and accuracy) of the resulting tree. Many popular split criteria suffer from bias towards predictor variables with large domains (White & Liu, 1994; Kononenko, 1995).

Consider two predictor variables X_1 and X_2 whose association with the class label is equally strong (or weak). Intuitively, a split selection criterion is unbiased if on a random instance the criterion chooses both X_1 and X_2 with probability $1/2$ as split variables. Unfortunately, this is usually not the case.

There are two previous meanings associated with the notion of *bias* in decision tree construction: an anomaly observed by Quinlan (1986), and the difference in distribution of the split criteria applied to different predictor variables (White & Liu, 1994). In this paper, we start in Section 3 by giving a precise, quantitative definition of bias in split variable selection. By extending the studies by White and Liu (1994) and Kononenko (1995), we quantify in an extensive experimental study the bias in split selection for the case that none of the predictor variables is correlated with the class label.

Section 4 contains the heart of our paper. Assume that we use split criterion $s(\mathcal{D}, X)$ to calculate the quality q of predictor variable X as split variable for training dataset $\mathcal{D}$. Consider the the p-value p of value q, which is the probability to see a value as extreme as the observed value q in the case that X is not correlated with the class label. In Section 4.1, we prove that choosing the variable with the lowest p-value results in a split selection criterion that is nearly unbiased — independent of the initial split criterion s. Since previous criteria such as χ^2 and G^2 (Mingers, 1987) and the permutation test (Frank & Witten, 1998) are p-values, our theorem explains why χ^2, G^2, and the permutation test are virtually unbiased. We continue in Section 4.2 by computing a tight approximation of the distribution of Breiman's `gini` index for k-ary splits which gives us a theoretical approximation of the p-value of the index. We demonstrate in Section 5 that our new criterion is nearly unbiased.

Note that the general method that we propose is similar in spirit but different from the work of Jensen and Cohen (2000) on the problems with multiple comparisons in induction algorithms. The bias in split selection for discrete variables is not due to multiple comparisons, but rather due to inherent statistical fluctuations as we explain in Section 3.

2. Preliminaries

In this section we introduce some notation and describe several popular split selection criteria.

2.1 Split Selection

Let $\mathcal{D}$ be the training dataset consisting of N datapoints. We consider without loss of generality the selection of the split variable at the root node of the

classification tree. Let X be a predictor variable, let $\{x_1, \ldots, x_n\}$ be the domain of X, and let N_i be the number of data-points in the dataset $\mathcal{D}$ for which $X = x_i$ for $i \in \{1, .., n\}$. Let $\{c_1, \ldots, c_k\}$ be the domain of the class label C, and let S_j be the number of training records in $\mathcal{D}$ for which $C = c_j$ for $j \in \{1, .., n\}$. Denote by A_{ij} the number of data-points for which $X = x_i \wedge C = c_j$. Also let p_j, $j \in \{1, .., k\}$ be the prior probability to see class label c_j in the dataset $\mathcal{D}$. Obviously the following normalization constraint holds: $\sum_{j=1}^{k} p_j = 1$.

Using the notation we just introduced, we can form a contingency table for dataset $\mathcal{D}$ as shown in Figure 1. We call the numbers on the last column and the last row *marginals* since they obey the following marginal constraints: $\sum_{i=1}^{n} N_i = N$, $\sum_{i=1}^{n} A_{ij} = S_j$, and $\sum_{j=1}^{k} A_{ij} = N_i$. Using the contingency table we have the following maximum likelihood estimates: $P[X = x_i] = N_i/N$, $p_j = P[C = c_j] = S_j/N$, $P[C = c_j \wedge X = x_i] = A_{ij}/N$ and $P[C = c_j | X = x_i] = A_{ij}/N_i$.

Note that this contingency table contains the sufficient statistics for split selection criteria that make univariate splits (Gehrke et al., 1998); thus given the table, any split selection criterion can compute the quality of X as split variable.

X	C_1	...	C_j	...	C_k	
x_1	A_{11}	...	A_{1j}	...	A_{1k}	N_1
.	.	...	.	...	.	.
.	.	...	.	...	.	.
x_i	A_{i1}	...	A_{ij}	...	A_{ik}	N_i
.	.	...	.	...	.	.
.	.	...	.	...	.	.
x_n	A_{n1}	...	A_{nj}	...	A_{nk}	N_n
	S_1	...	S_j	...	S_k	N

Figure 1. Contingency table for a generic dataset $\mathcal{D}$ and generic predictor variable X.

2.2 Split selection criteria

Let us briefly define some popular split criteria using the maximum likelihood estimates of probabilities described in the previous section.

χ^2 Statistic. Used in (Mingers, 1987):

$$\chi^2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(A_{ij} - E(A_{ij}))^2}{E(A_{ij})}, \quad E(A_{ij}) = \frac{N_i S_j}{N} \quad (1)$$

Gini Gain. Defined in (Breiman et al., 1984) as:

$$\begin{aligned} \Delta g &\stackrel{\text{def}}{=} \sum_{i=1}^{n} P[X = x_i] \sum_{j=1}^{k} P[C = c_j | X = x_i]^2 \\ &\quad - \sum_{j=1}^{k} P[C = c_j]^2 \\ &= \frac{1}{N} \sum_{j=1}^{k} \left(\sum_{i=1}^{n} \frac{A_{ij}^2}{N_i} - \frac{S_j}{N} \right) \end{aligned} \quad (2)$$

Information Gain. Defined in (Quinlan, 1986) as:

$$\begin{aligned} IG &\stackrel{\text{def}}{=} \sum_{j=1}^{k} \Phi(P[C = c_j]) + \sum_{i=1}^{n} \Phi(P[X = x_i]) \\ &\quad - \sum_{j=1}^{k} \sum_{i=1}^{n} \Phi(P[C = c_j \wedge X = x_i]) \\ &= \frac{1}{N} \left(\sum_{j=1}^{k} \sum_{i=1}^{n} A_{ij} \log A_{ij} - \sum_{j=1}^{k} S_j \log S_j \right. \\ &\quad \left. - \sum_{i=1}^{n} N_i \log N_i + N \log N \right), \end{aligned} \quad (3)$$

where $\Phi(p) = -p \log p$

Gain Ratio. Defined in (Quinlan, 1986) as:

$$\begin{aligned} GR &\stackrel{\text{def}}{=} \frac{IG}{\sum_{i=1}^{n} \Phi(P[X = x_i])} \\ &= \frac{IG}{\frac{1}{N}(N \log N - \sum_{i=1}^{n} N_i \log N_i)} \end{aligned} \quad (4)$$

G^2 Statistic. Used in (Mingers, 1987):

$$G^2 \stackrel{\text{def}}{=} 2 \cdot N \cdot IG \log_e 2 \quad (5)$$

3. Bias in Split Selection

In this section we introduce formally the notion of bias in split variable selection for the case that there is no correlation between predictor variables and the class label (i.e., the predictor variables are unpredictive of the class label). We then show that three popular split selection criteria are biased towards predictor variables with large domains.

3.1 A Definition of Bias

In order to study the behavior of the split criteria for the case where there is no correlation between a predictor variable and the class label we formalize the following setting:

Null Hypothesis: For every $i \in \{1,..,n\}$, the random vector $(A_{i1},\ldots,A_{ik})$ has the distribution Multinomial$(N_i, p_1,\ldots,p_k)$.

Intuitively, the Null Hypothesis assumes that for each value of the predictor variable, the distribution of the class label results from pure multi-face coin tossing, thus the distribution of the class label obeys a multinomial distribution. Since $\sum_{i=1}^{n} A_{ij} = S_j$, the random vector $(S_1,\ldots,S_k)$ has the distribution Multinomial$(N, p_1,\ldots,p_k)$.

We now give a formal definition of the bias. Let s be a split criterion, and let $s(\mathcal{D}, X)$ be the value of s when applied to dataset $\mathcal{D}$. Usually the split variable selection method compares the values of the split criteria for two variables and picks the one with the biggest corresponding value for predictor attribute X.[1] Now let $\mathcal{D}$ be a random dataset whose values are distributed according to the Null Hypothesis. Thus $s(\mathcal{D}, X)$ is now a random variable that has a given distribution under the Null Hypothesis. Define the probability that split selection method s chooses predictor variable X_1 over X_2 as follows:

$$P_s(X_1, X_2) \stackrel{\text{def}}{=} P[s(\mathcal{D}, X_1) > s(\mathcal{D}, X_2)]. \qquad (6)$$

We can now define the bias of the split criterion between X_1 and X_2 as the logarithmic odds of choosing X_1 over X_2 as a split variable when neither X_1 nor X_2 is correlated with the class label, formally:

$$\text{Bias}(X_1, X_2) = \log_{10}\left(\frac{P_s(X_1, X_2)}{1 - P_s(X_1, X_2)}\right). \qquad (7)$$

When the split criterion is unbiased, then $Bias(X_1, X_2) = \log_{10}(0.5/(1-0.5)) = 0$. The bias is positive if s prefers X_1 over X_2 and negative, otherwise. Bigger value of $|Bias(X_1, X_2)|$ indicate stronger bias, and we desire split criteria with values of the bias as close to 0 as possible.

Our notion of bias is inherently statistical in nature, and it reflects the intuition that under the Null Hypothesis the split criterion should have no preference for any predictor variable. There have been several attempts to define bias in split variable selection. Quinlan's Gain Ratio (Quinlan, 1986) was designed to correct for an anomaly that he observed, but as we will show in Section 3.2, the Gain Ratio merely reduces the bias, but it does not remove it. White and Liu (1994) point out that Quinlan's definition of the bias is non-statistical in nature; their definition of the bias is based on the distribution of the split criterion for predictor variables with the same number of values. It is harder to use in practice since it implies a test of the equality of two distributions instead of two numbers as in our case. Loh and Shih (1997) introduce a notion of bias whose formalization coincides with our definition.

[1] For the case when smaller values of the split criterion are preferable, we can use $-s$ as split criterion.

3.2 Experimental Demonstration of the Bias

We performed an extensive experimental study to demonstrate the bias according to our definition in Section 3.1. We generated synthetic training datasets with two predictor variables and two class labels. We chose $n_1 = 10$ different variable values for predictor variable X_1 and $n_2 = 2$ different variable values for predictor variable X_2.[2] We varied N, the size of the training database from 10 and 1000 records in steps of 40 records, and we varied the value of the prior probability p_1 of the first class label exponentially between 0 and 1/2. Since all split criteria are invariant to class labels permutations, the graphs depicting the bias are symmetric with respect to $p_1 = 1/2$; we present here the part of the graphs with $p_1 \leq 1/2$. To estimate $P_s(X_1, X_2)$, we performed 100000 Monte Carlo trials in which we generated random training databases distributed according to the Null Hypothesis (thus the standard error of all our measurements is smaller than 0.0016). Exactly the same random instances were used for all split criteria.

The results of our experiments are shown in Figures 2 to 6. Figure 2 shows the bias of the `gini` gain, Figure 3 shows the bias of the information gain, Figure 4 shows the bias of Quinlan's gain ratio, Figure 5 shows the bias of the p-value of the χ^2-test according to the χ^2 distribution (with $n-1$ degrees of freedom), and Figure 6 shows the bias of the p-value of the G^2-statistics according to the χ^2 distribution (with $n-1$ degrees of freedom). The χ^2-distribution with $n-1$ degrees of freedom has to be used since there are $2n$ entries in the contingency table with n marginal constraints ($\sum_{j=1}^{k} A_{ij} = N_i$) and the additional constraint that S_j/N is used as an estimate for p_j.

For values of p_1 between 10^{-2} and 1/2 both the `gini` gain and the information gain show a very strong bias (X_1 is chosen $10^{1.80} = 63$ times more often than X_2), the gain ratio is less biased (X_1 is chosen $10^{0.8} = 6.3$ times more often than X_2), but the bias is still significant. The χ^2 test is basically unbiased in this region except for really small values of N. The G^2 test is unbiased for large values of N and for p_1 close to 1/2, but

[2] Due to space limitation, we cannot present the full scope of experiments we performed. Results from experiments with different values for n_1 and n_2 were qualitatively similar.

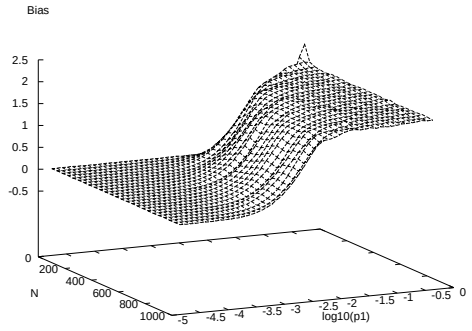

Figure 2. The bias of the `gini` gain.

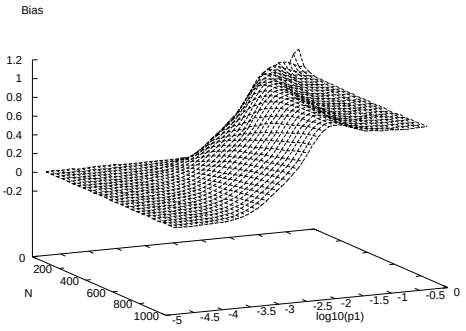

Figure 4. The bias of the gain ratio.

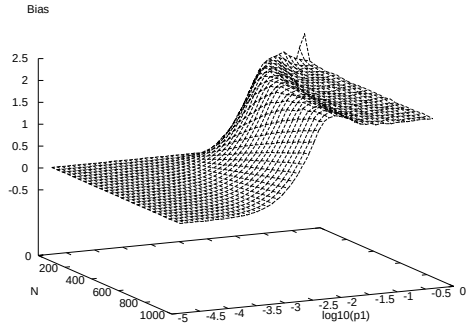

Figure 3. The bias of the information gain.

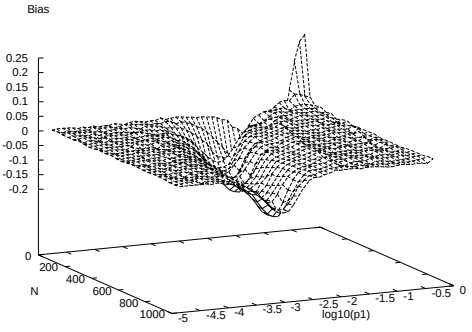

Figure 5. The bias of the p-value of the χ^2-test (using a χ^2-distribution).

the bias is noticeable in important border cases that are relevant in practice (for example for $p_1 = 10^{-2}$ and $N = 1000$, the bias has value 0.20).

For values of p_1 between 10^{-4} and 10^{-2}, the `gini` gain, the information gain, and the gain ratio start having less and less bias. Both the χ^2 and the G^2 criterion have a preference towards variable variables with few values, the bias gets as low as -0.2 (corresponds to 1.58 odds) when $p_1 N = 1$. The maximum negative bias corresponds to datasets that on average have a single data-point with class label c_1. We postpone the explanation of this phenomenon to Section 5. The region where $p_1 < 10^{-4}$ corresponds to training datasets where no record has class label c_1 (all records have the same class label). In this case the `gini` gain, the information gain, and the gain ratio have value 0, whereas the χ^2 and G^2 criteria have value 1, independent of the split variable. In our experiments, we tossed a fair coin in case that the split criterion returns the same value for variables X_1 and X_2, thus the bias is basically 0.

One surprising insight from our experiments is that the bias for the `gini` gain, the information gain, and the gain ratio do not vanish as N gets arbitrary large. In addition, the bias does not seem to have a significant dependency on p_1 as long as all entries in the contingency table for variable X_1 are moderately populated.

We obtained similar results for different variable domain sizes. The bias is more pronounced for bigger differences in the domain sizes of X_1 and X_2. When the domain sizes are identical ($n_1 = n_2$), the bias is almost nonexistent. These facts suggest that the size of the domain is the most significant factor that influences the behavior under the Null Hypothesis. This conclusion, for the `gini` gain, is supported by the theoretical formulas in Section 4.2.

The bias for the `gini` gain, the information gain, and the gain ratio comes from the fact that under the Null Hypothesis the value of the split criterion is not exactly zero. The values of $s(X, \mathcal{D})$ monotonically increase with n, the size of the domain of X, and variables with

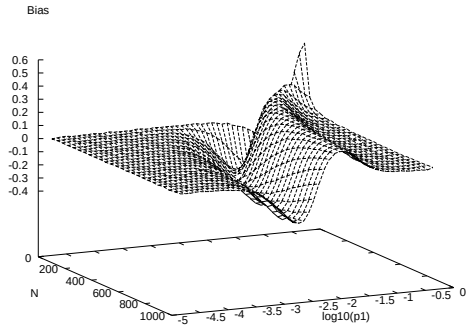

Figure 6. The bias of the p-value of the G^2-test (using a χ^2-distribution).

more values tend to have larger values of $s(X, \mathcal{D})$ due to the fact that the counts in the contingency table have bigger statistical fluctuations. The bias is thus due to the inability of traditional split criteria to account for these normal statistical fluctuations. In the next section, we will present a technique that allows us to remove the bias from existing split criteria.

4. Correction of the Bias

In Section 4.1, we present a general method for removing the bias of any arbitrary split criterion. We then show in Section 4.2 how our method can be used to correct the bias inherent in the `gini` gain.

4.1 A General Method for Bias Removal

Let us first give some intuition behind our method. We observed in Section 3.2 that the expected value of several split criteria under the Null Hypothesis depends on the size of the domain of the predictor variables. Assume that the value of the split criterion for variable X_1 is v_1 (X_2 and v_2, respectively). Instead of comparing v_1 and v_2 directly (and incurring a biased variable selection), we compute the p-value p_1, the probability that the value of the split criterion is as extreme as v_1 under the Null Hypothesis (and we compute p_2, analogously). We then choose the split variable with the lower p-value. The remainder of this section is devoted to a formal proof that the p-value of any split criterion is virtually unbiased under the assumption that the Null Hypothesis holds.

Let X and X_H be two identically distributed random variables (i.e., $\forall x \in \mathrm{Dom}(X) : p_x \stackrel{\text{def}}{=} P[X = x] = P[X_H = x]$), and let Y and Y_H be two other identically distributed random variables. Define $C_X(x) \stackrel{\text{def}}{=} 1 - P[X_H \leq x] = 1 - \sum_{x' \leq x} p_{x'}$, and define $C_Y(y)$ analogously. Let $\Delta \stackrel{\text{def}}{=} \max_x P[X = x] + \max_y P[Y = y]$.

Lemma 1 *Let X and Y be two independent discrete random variables. Then $\forall \gamma \in [0,1]$:$P[C_X(X) < C_Y(Y)] + \gamma P[C_X(X) = C_Y(Y)] \in (1/2 - \Delta, 1/2 + \Delta)$.*

Proof:

$$\begin{aligned}\mathcal{P} &= P[C_X(X) < C_Y(Y)] \\ &= \sum_x \sum_y I(C_X(x) < C_Y(y)) P[X = x \wedge Y = y] \\ &= \sum_x \sum_y I\left(\sum_{x'}^{x} p_{x'} > \sum_{y'}^{y} p_{y'}\right) p_x p_y\end{aligned} \tag{8}$$

where $I(\cdot)$ is the indicator function.

For a fixed x, let y_x be the biggest value of $y \in \mathrm{Dom}(Y)$ such that $\sum_{x'<x} p_{x'} > \sum_{y'<y} p_{y'}$ still holds. Equation 8 then can be rewritten as follows:

$$\mathcal{P} = \sum_x p_x \sum_y^{y_x} p_y \tag{9}$$

On the other hand using the definition of y_x we have:

$$\sum_{x'}^{x} p_{x'} - \sum_{y}^{y_x} p_y > 0, \text{and} \tag{10}$$

$$\sum_{x'}^{x} p_{x'} - \sum_{y}^{y_x} p_y \leq p_{y_x^+} \leq \max_y p_y \tag{11}$$

where $p_{y_x^+}$ is the smallest $y \in \mathrm{Dom}(Y)$ such that $y > y_x$. The previous two inequalities imply:

$$\sum_{x'}^{x} p_{x'} - \max_y p_y \leq \sum_{y}^{y_x} p_y < \sum_{x'}^{x} p_{x'} \tag{12}$$

Multiplying by p_x, summing up on x and using the result of Equation 9 we obtain:

$$\sum_x p_x \sum_{x'}^{x} p_{x'} - \max_y p_y \leq \mathcal{P} < \sum_x p_x \sum_{x'}^{x} p_{x'} \tag{13}$$

To further simplify, let X' be a random variable with the same distribution as X. We then obtain:

$$\begin{aligned}\sum_x p_x \sum_{x'}^{x} p_{x'} &= \sum_x I(x' \leq x) P[X' = x' \wedge X = x] \\ &= P[X' \leq X] = \frac{1}{2} - \frac{1}{2} P[X' = X] \\ &= \frac{1}{2} - \frac{1}{2} \sum_x p_x^2 \\ &\in \left(\frac{1}{2} - \frac{1}{2} \max_x p_x, \frac{1}{2} - \frac{1}{2} \min_x p_x\right)\end{aligned} \tag{14}$$

Using Equations 13 and 14 we get:

$$\frac{1}{2} - \max_x p_x - \max_y p_y < \mathcal{P} < \frac{1}{2} - \frac{1}{2}\min_x p_x \quad (15)$$

If the roles of x and y are switched we obtain:

$$\frac{1}{2} - \Delta < P[C_X(X) > C_Y(Y)] < \frac{1}{2} - \frac{1}{2}\min_y p_y, \quad (16)$$

which implies:

$$\frac{1}{2} + \frac{1}{2}\min_y p_y < P[C_X(X) \leq C_Y(Y)] < \frac{1}{2} + \Delta, \quad (17)$$

thus

$$\frac{1}{2} - \Delta < \mathcal{P} + \gamma P[C_X(X) = C_Y(Y)] < \frac{1}{2} + \Delta \quad (18)$$

□

According to Lemma 1, if the p-value of a criterion is used to decide the split variable, the probability of choosing one variable over another is not farther than Δ from $\frac{1}{2}$. In practice, even for small sizes of the dataset, any split criterion has a huge number of possible values and the probability of the criterion to take on a value is much smaller than $\frac{1}{2}$, thus $\Delta \approx 0$, and the p-value is a virtually unbiased split criterion. Thus we can guarantee that the p-value of any split criterion s is unbiased under the Null Hypothesis, as long as s does not take on a single value with a significantly large probability.

Using the above fact, a general method to remove the bias in split variable selection consists of two steps. First, we compute the value v of the original split criteria s on the given dataset. Second, we compute the p-value of v under the Null Hypothesis and we select the variable with the smallest p-value as the split variable.

The above method requires the computation of the p-value of a given criterion. We can distinguish four ways in which this can be accomplished.

- Exact computation. Use the exact distribution of the split criterion. The main drawback is that this is almost always very expensive; it is reasonably efficient only for $n = 2$ and $k = 2$ (Martin, 1997).
- Bootstrapping. Use Monte Carlo simulations with random instances generated according to the Null Hypothesis. This method was used in by Frank and Witten (1998); its main drawback is the high cost of the the Monte Carlo simulations.
- Asymptotic approximations. Use an asymptotic approximation of the distribution of the split criterion (e.g., use the χ^2-distribution to approximate the χ^2-test (Kass, 1980) and the G^2-statistic (Mingers, 1987)). Approximations often work well in practice, but they can be inaccurate for border conditions (e.g., small entries in the contingency table).
- Tight approximations. Use a tight approximation of the distribution of the criterion with a nice distribution. While conceptually superior to the previous three methods, such tight approximations might be hard to find.

4.2 A Tight Approximation of the Gini Gain

In this section we give a tight approximation of the distribution of the `gini` gain, and we use our approximation in combination with Lemma 1 to compose a new unbiased split criterion.

Note that the p-value of the `gini` gain can be well approximated if the cumulative distribution function (c.d.f) of the distribution of the `gini` gain can be well approximated (since the p-value=1-c.d.f.). We experimentally observed by looking at the shape of the probability distribution function of the `gini` gain that it is very close to the shape of distributions from the gamma family.[3] Our experiments show that the Gamma distribution – using the expected value and variance of the `gini` gain as distribution parameters (which completely specify a gamma distribution) – is a very good approximation of the distribution of the `gini` gain. In the remainder of this section, we will show how to compute exactly the expected value and the variance of the `gini` gain under the Null Hypothesis, and we will use these values for a tight approximation of the `gini` gain with the Gamma distribution.

As mentioned in Section 2, the contingency table described in Section 2 contains the sufficient statistics for the computation of the `gini` gain. Thus in order to analyze the distribution of the `gini` gain, it is sufficient to look at the distribution of the entries in the contingency table. Consider a given fixed set of parameters N, n, k, $N_i, i \in \{1, .., n\}$, and p_j, $j \in \{1, .., k\}$. If the Null Hypothesis holds, the A_{ij}'s and S_j's are random variables with multinomial distributions (see Section 3). Using the definition of the `gini` gain (Equation 2), linearity of expectation, the fact that the A_{ij}'s and S_j's have multinomial distributions, and the normalization constraint on the p_j's, we get the following

[3] Due to limited space, we have omitted results from these experiments in this paper.

formula for the expectation $E(\Delta g)$ of the `gini` gain under the Null Hypothesis:

$$\begin{aligned} E(\Delta g) &= \frac{1}{N}\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\frac{E(A_{ij}^2)}{N_i} - \frac{E(S_j)}{N}\right) \\ &= \frac{1}{N}\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\frac{N_i p_j(1-p_j+N_i p_j)}{N_i} \right. \\ &\qquad \left. - \frac{N p_j(1-p_j+N p_j)}{N}\right) \\ &= \frac{1}{N}\sum_{j=1}^{k}(n p_j(1-p_j) + N p_j^2 \\ &\qquad - p_j(1-p_j) - N p_j^2) \\ &= \frac{n-1}{N}\left(1-\sum_{j=1}^{k}p_j^2\right) \end{aligned} \tag{19}$$

so the expected value of the `gini` gain is indeed linear in n as observed by White and Liu (1994).

Computation of the the variance $\mathrm{Var}\,(\Delta g)$ of the `gini` gain results in the following formula:[4]

$$\begin{aligned} &\mathrm{Var}\,(\Delta g) = \\ &\frac{1}{N^2}\left[(n-1)\left(2\sum_{j=1}^{k}p_j^2 + 2\left(\sum_{j=1}^{k}p_j^2\right)^2 - 4\sum_{j=1}^{k}p_j^3\right)\right. \\ &+\left(\sum_{i=1}^{n}\frac{1}{N_i} - 2\frac{n}{N} + \frac{1}{N}\right)\times \\ &\left.\left(-2\sum_{j=1}^{k}p_j^2 - 6\left(\sum_{j=1}^{k}p_j^2\right)^2 + 8\sum_{j=1}^{k}p_j^3\right)\right] \end{aligned} \tag{20}$$

Note that our formulas for the expected value and the variance of the `gini` gain under the Null Hypothesis are not approximations, but exact values. To find the right parameters of the suitable gamma distribution with the same expected value and same variance as the `gini` gain, we use the fact that $E(\Gamma(\alpha,\theta)) = \alpha\theta$ and $\mathrm{Var}\,(\Gamma(\alpha,\theta)) = \alpha\theta^2$, thus $\alpha = E(\Delta g)^2/\mathrm{Var}\,(\Delta g)$ and $\theta = \mathrm{Var}\,(\Delta g)/E(\Delta g)$. Approximating the p-value of `gini` gain with the p-value of this distribution we obtain:

$$\text{p-value}(\Delta g_e) = 1 - Q\left(\frac{E(\Delta g)^2}{\mathrm{Var}\,(\Delta g)}, \frac{\Delta g_e \mathrm{Var}\,(\Delta g)}{E(\Delta g)}\right), \tag{21}$$

where Δg_e is the actual value for `gini` computed on the given dataset and $Q(x,y) = \Gamma(x,y)/\Gamma(x)$ is the regularized incomplete gamma function. We call this new criterion the *Gamma correction*. From Equations 19, 20 and 21 it is easy to see that the correction depends only on n, N, and the N_i's and p_j's.

Note that there is a very important numerical precision problem associated with the above formula. Even for moderate correlation between a predictor variable and the class label, the value of the second term in Equation 21 approaches 1 very rapidly (by far exceeding the precision of the processor). Thus the computed value of the p-value is 0 in this case, seemingly limiting the usefulness of our criterion for the case that correlations between a predictor variable and the class label are present. This "non-discrimination" anomaly was also observed by Kononenko (1995).

For our criterion, we can avoid this problem by directly computing the logarithm of the p-value using a series expansion.[5] In this fashion, values of the logarithm of the p-value (which can be used instead of the p-value since the logarithm is a monotonically increasing function) can be computed accurately even for datasets with millions of records and very strong correlations.

The computational complexity of our new criterion is $O(n+k)$ since we have to compute the sum of inverses of the N_i's and p_j's; all other factors can be computed in time $O(1)$, including the logarithm of the incomplete regularized gamma function. Thus our new criterion can be computed very efficiently in practice.[6]

5. Experimental Evaluation

In this section we will show experimental evidence that our theoretical corrections behave well in practice. To evaluate the bias of the gamma correction of the `gini` gain we repeated the experiment from Section 3. The bias of our correction of the `gini` gain as a function of N and p_1 is depicted in Figure 7. As can be observed by comparing Figures 5 and 7, the bias for the corrected `gini` gain and the χ^2-test are practically identical for all values of p_1 and N. Also, for p_1 between 10^{-4} and 10^{-2} all the statistical methods are biased towards predictor variables with small n in precisely

[4] Due to space constraints we omitted the proof of this result.

[5] We used the implementation of the incomplete gamma function in the Statistics package ANA (Shine & Strous, 2001)

[6] On a Pentium III 933MHz the computation of the incomplete regularized gamma function takes $155\mu s$. This is also the time to compute the contingency table for 14000 samples in the most favorable case (one predictor variable and highly optimized code for this special case).

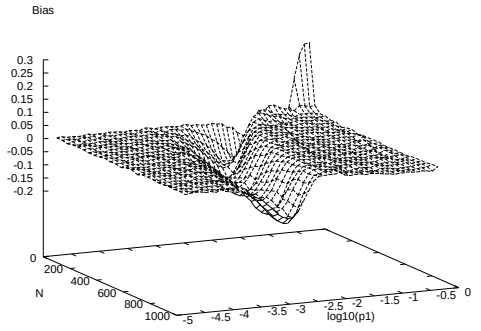

Figure 7. Bias of the p-value of the `gini` gain using the gamma correction

the same way. As mentioned in Section 3, the most extreme bias is obtained for $p_1 = 1/N$. In this case the probability to see exactly one data-point with class label c_1 is $N\frac{1}{N}\left(1-\frac{1}{N}\right)^{N-1} \approx e^{-1}$. The margin Δ from the Lemma in Section 4 is at least $2e^{-1} \approx 0.73$ which means that the exact correction (using the exact distribution of the split criteria) can have any bias. Thus around $p_1 = 1/N$ we cannot expect any of the statistical methods to be perfectly unbiased.

Note that for small entries in the contingency table the χ^2-distribution is a poor approximation of the χ^2-test. (We observed that for this case the expected value according to the χ^2-distribution is correct, but the variance is overestimated.) In the case that a predictor variable is not correlated with the class label, the overestimation of the variance does not seem to matter (but this might not be the case when correlations are present).

To summarize our experiments, the gamma correction of the `gini` gain and the χ^2 criterion have very good behavior under the Null Hypothesis. The G^2 criterion behaves well if class labels are almost equiprobable but some bias is present if this is not the case. The `gini` gain, the information gain, and the gain ratio have significant biases towards variables with more values.

6. Conclusions

This paper addresses the fundamental problem of bias in split variable selection in classification tree construction. Our contribution is (1) a general method to provably remove the bias introduced by categorical variables with large domains and (2) an the application of our method to the removal of the bias for the `gini` gain.

Previous work for some split criterions suggests that removal of the bias by the usage of p-values improves the quality of the split when correlations are weak and in the same time preserves the good behavior for strong correlations (Mingers, 1987; Frank & Witten, 1998). This suggests that bias removal in general is useful in practice. We consider the work described in this paper an initial step in a potentially interesting direction. In future work we intend to thoroughly investigate, both theoretically and experimentally, the properties of the proposed correction of the split criteria for the case when correlations are present.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Belmont: Wadsworth.

Frank, E., & Witten, I. H. (1998). Using a permutation test for attribute selection in decision trees. *International Conference on Machine Learning.*

Gehrke, J., Ramakrishnan, R., & Ganti, V. (1998). Rainforest – a framework for fast decision tree construction of large datasets. *Proceedings of the 24th VLDB Conference* (pp. 416–427).

Jensen, D. D., & Cohen, P. R. (2000). Multiple comparisons in induction algorithms. *Machine Learning, 38*, 309–338.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 119–127.

Kononenko, I. (1995). On biases in estimating multi-valued attributes.

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica, 7.*

Martin, J. K. (1997). An exact probability metric for decision tree splitting. *Machine Learning, 28*, 257–291.

Mingers, J. (1987). Expert systems – rule induction with statistical data. *J. Opl. Res. Soc., 38*, 39–47.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.

Shine, R. A., & Strous, L. (2001). Ana. http://ana.lmsal.com.

White, A. P., & Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Machine Learning, 15*, 321–329.