Using a Permutation Test for Attribute Selection in Decision Trees

Eibe Frank
Department of Computer Science
University of Waikato
Hamilton, New Zealand
eibe@cs.waikato.ac.nz

Ian H. Witten
Department of Computer Science
University of Waikato
Hamilton, New Zealand
ihw@cs.waikato.ac.nz

Abstract

Most techniques for attribute selection in decision trees are biased towards attributes with many values, and several *ad hoc* solutions to this problem have appeared in the machine learning literature. Statistical tests for the existence of an association with a prespecified significance level provide a well-founded basis for addressing the problem. However, many statistical tests are computed from a chi-squared distribution, which is only a valid approximation to the actual distribution in the large-sample case—and this patently does not hold near the leaves of a decision tree. An exception is the class of permutation tests. We describe how permutation tests can be applied to this problem. We choose one such test for further exploration, and give a novel two-stage method for applying it to select attributes in a decision tree. Results on practical datasets compare favorably with other methods that also adopt a pre-pruning strategy.

1 Introduction

Statistical tests provide a set of theoretically well-founded tools for testing hypotheses about relationships in a set of data. One pertinent hypothesis, when selecting attributes for a decision tree, is whether there is a significant association between an attribute's values and the classes. With r attribute values and c classes, this equates to testing for independence in the corresponding $r \times c$ contingency table (White & Liu, 1994), and statistical tests designed for this purpose can be applied directly. Unlike most commonly-used attribute selection criteria, such tests are not biased towards attributes with many values, which is important because it prevents the decision tree induction algorithm from selecting splits that overfit the training data by being too fine-grained.

Statistical tests are based on probabilities derived from the distribution of a test statistic. Two popular test statistics for assessing independence in a contingency table have been proposed for attribute selection: the chi-squared statistic χ^2 and the log likelihood ratio G_2 (White & Liu, 1994). For large samples, both are distributed according to the chi-squared distribution. But this is not the case for small samples (Agresti, 1990)—and small samples inevitably occur close to the leaves in a decision tree. Thus it is inadvisable to use probabilities derived using the chi-squared distribution for decision tree induction.

Fortunately, there is an alternative that does apply in small frequency domains. In statistical tests known as "permutation tests" (Good, 1994), the distribution of the statistic of interest is calculated directly instead of relying on the chi-squared approximation—in other words they are "non-parametric" rather than "parametric." Such tests do not suffer from the small expected frequency problem because they do not use the chi-squared approximation.

This paper describes the application of permutation tests to attribute selection in a decision tree. We examine one such test—the Freeman and Halton test—in detail by performing experiments on artificial and practical datasets: the results show that this method is indeed preferable to a test that assumes the chi-squared distribution. The statistic of the Freeman and Halton test is the exact probability p_f of a contingency table f given its marginal totals (Good, 1994). Recently, Martin (1997) investigated the use of this statistic, p_f, directly for attribute selection. We show

that results can be improved by using it in conjunction with the Freeman and Halton test.

Section 2 introduces the idea of permutation tests and how they can be used to test significance in a contingency table. In Section 2.2 we describe the Freeman and Halton test. The test is expensive, but simple computational economies are described in Section 2.3. Section 2.4 describes a novel two-stage method, based on these ideas, for selecting attributes in a decision tree. Section 3 presents experimental results on artificial and standard datasets. We verify that the Freeman and Halton test does not prefer attributes with many values, whereas the test statistic p_f by itself is biased. We also verify that the parametric version of the chi-squared test is biased in small-frequency domains. Finally, we demonstrate that good results are obtained when the new method is applied to decision-tree building. Section 4 reviews existing work on using statistical tests for contingency tables in machine learning, while Section 5 contains some concluding remarks.

2 A Permutation Test and its Application to Attribute Selection

The procedure for permutation tests is simple (Good, 1994). First, a test statistic is chosen that measures the strength of the effect being investigated, and is computed over the data. The null hypothesis is that the observed strength of the effect is *not* significant. Next, the labels of the original data are permuted and the same statistic is calculated for the relabeled data; this is repeated for all possible permutations of labels. The idea is to ascertain the likelihood of an effect of the same or greater strength being observed fortuitously on randomly labeled data with identical marginal properties. Third, the test statistic's value for the original data is compared with the values obtained over all permutations, by calculating the percentage of the latter that are at least as extreme, or more extreme, than the former. This percentage constitutes the significance level at which the null hypothesis can be rejected, in other words, the level at which the observed strength of the effect can be considered significant.

2.1 Permutation Tests for Contingency Tables

Contingency tables summarize the observed relationship between two categorical response variables. Several different statistics can be used to measure the strength of the dependency between two variables (Good, 1994), the two most common being the chi-squared statistic χ^2 and the log likelihood ratio G_2. The standard tests using these statistics are based on the fact that the sampling distribution of both statistics is well-approximated by the chi-squared distribution. They calculate the significance level directly from that distribution.

Unfortunately, as noted in the introduction, the chi-squared distribution assumption is only valid for either statistic when the sample size is large enough. The chi-squared distribution approximates the true sampling distribution poorly if the sample size is small (or the samples are distributed unevenly in the contingency table). In a decision tree the sample size becomes smaller and smaller and the distribution of the samples more and more skewed the closer one gets to the leaves of the tree. Thus one cannot justify using a test based on the chi-squared approximation for significance testing throughout a decision tree (although one might at the upper levels where samples are large). Permutation tests offer a theoretically sound alternative that is admissible for any sample size.

The standard permutation test for $r \times c$ contingency tables, which we have also chosen to employ for this paper, is based on the statistic p_f, the exact probability of a contingency table given its marginal totals. It is known as the "Freeman and Halton" test and it is a generalization of Fisher's exact test for 2×2 tables (Good, 1994). However, we emphasize that other test statistics could equally well be used, thereby obtaining exact, non-parametric, versions of conventional parametric tests that are valid in small-frequency domains (Good, 1994).[1]

2.2 Testing the Significance of an Attribute

For attribute selection, we seek to test whether there is a significant association between an attribute's values and the class values. With r attribute values and c classes, this is the same as testing for independence in the corresponding $r \times c$ contingency table (White & Liu, 1994).

If the $r \times c$ contingency table f contains the frequencies f_{ij} with column marginals $f_{.j}$ and row marginals $f_{i.}$, the probability p_f of this table is given by

[1] We have also used a permutation test based on χ^2, instead of on p_f, in all the experiments described in Section 3, and obtained almost identical results.

$$p_f = \frac{\prod_{i=1}^{r} f_{i.}! \prod_{j=1}^{c} f_{.j}!}{f_{..}! \prod_{i=1}^{r} \prod_{j=1}^{c} f_{ij}!}.$$

Permuting the instances' class labels does not affect the row and column totals, and therefore the set of all permutations of the class labels corresponds to the set of all contingency tables with the same row and column totals. If p is the proportion of tables for which p_f is less than or equal to the probability p_o of the original table, then

$$p = \sum I(p_f \leq p_o) p_f,$$

where $I(.)$ denotes the indicator function, constitutes the p-value of the Freeman and Halton test. The function computing p is known as a multiple hypergeometric distribution (Agresti, 1990). The resulting value of p is simply compared with a prespecified desired significance level.

2.3 Approximating the Exact Test

Exact computation of the p-value of a permutation test is only possible for sparsely populated tables, and is computationally infeasible for most tables resulting from practical machine learning datasets. Fortunately, p can be approximated to arbitrary precision by Monte Carlo sampling as follows (Good, 1994).

For each of n trials the class labels are randomly permuted, the test statistic is computed, and its value is compared to the value for the original (unpermuted) data. The percentage of trials for which the artificially generated value is less than or equal to the original value constitutes an estimate $\hat{p}$ of the exact significance level p. This estimate is a binomial random variable with standard error $se(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$, and so its $100(1-\alpha)\%$ confidence interval is $\hat{p} \pm t_{n-1}(\alpha/2) se(\hat{p})$, where $t_{n-1}(\alpha/2)$ is obtained from Student's t-distribution.

This information is used to decide when to stop performing trials. Let p_{fixed} be the prespecified desired minimum significance level that an attribute must achieve unless it is to be considered independent of the class—the level at which the null hypothesis of "no significant dependence" is to be rejected. Then, with probability $(1-\alpha)$,

$$p > p_{\text{fixed}} \quad \text{if} \quad p_{\text{fixed}} < \hat{p} - t_{n-1}(\alpha) se(\hat{p}),$$

and

$$p < p_{\text{fixed}} \quad \text{if} \quad p_{\text{fixed}} > \hat{p} + t_{n-1}(\alpha) se(\hat{p})).$$

If the first inequality holds we judge the attribute to be significant; if the second holds we do not.[2] As n increases, the likelihood that one of the two inequalities will be true increases, but if p is very close to p_{fixed}, neither inequality will become true in a reasonable amount of time. Therefore the procedure is terminated when the number of trials reaches a prespecified maximum,[3] and any attribute that survives this number of trials is considered significant. The introduction of this cut-off point slightly increases the probability that an attribute is incorrectly judged to be significant.

2.4 Procedure for Attribute Selection

At each node of a decision tree we must decide which attribute to split on. This is done in two steps. First, attributes are rejected if they show no significant association to the class according to a pre-specified significance level. To judge "significance" we employ the Freeman and Halton test, approximated by Monte Carlo sampling as described above. Second, from the attributes that remain, the one with the lowest value of p_f is chosen.[4] The selected attribute is then used to split the set of instances, and the algorithm recurses.

The division into two steps is a crucial part of the procedure. It distinguishes clearly between the different concepts of *significance* and *strength*. For example, it is well known that the association between two distributions may be very significant even if that association is weak—if the quantity of data is large enough (Press, Teukolsky, Vettering & Flannery, 1988, p. 628). First, we test the significance of an association using a permutation test (specifically, the Freeman and Halton test); then we consider its strength (as measured by the exact probability p_f).

If no significant attributes are found in the first step, the splitting process stops and the subtree is not expanded any further. This gives an elegant, uniform, technique for pre-pruning.

3 Experimental Results

We begin with two controlled experiments that are designed to verify the relative performance of (a) the use

[2] Here, α is used instead of $\alpha/2$ because the comparisons are one-sided. In our experiments we set α to 0.005.

[3] We use at least 100 and at most 1000 trials in our experiments.

[4] Other attribute selection criteria could be employed at this stage; p_f was chosen to allow a direct comparison with the method proposed by Martin (1997).

Table 1: Average probabilities for random data (600 instances; uniformly distributed attribute values)

Attribute Values	Class Values	(a) $\hat{p}$	(b) p_f	(c) p_χ
2	2	0.525	0.045	0.488
2	5	0.511	1.63e-05	0.509
2	10	0.506	1.80e-10	0.505
5	2	0.497	1.55e-05	0.496
5	5	0.500	9.25e-18	0.497
5	10	0.491	6.62e-35	0.487
10	2	0.498	1.77e-10	0.495
10	5	0.520	7.84e-35	0.515
10	10	0.512	4.89e-68	0.503

Table 2: Average probabilities for random data (20 instances; non-uniformly distributed attribute values)

Attribute Values	Class Values	(a) $\hat{p}$	(b) p_f	(c) p_χ
2	2	0.745	0.285	0.515
2	5	0.674	0.024	0.466
2	10	0.741	0.004	0.446
5	2	0.549	0.027	0.444
5	5	0.561	1.02e-4	0.448
5	10	0.632	1.80e-6	0.418
10	2	0.548	0.004	0.430
10	5	0.581	1.72e-6	0.425
10	10	0.639	1.42e-8	0.382

of the exact-probability p_f statistic in the Freeman and Halton test, (b) the use of p_f by itself with no significance test (Martin, 1997), and (c) the use of the parametric version of the chi-squared test, that is, the probability of χ^2 calculated from the chi-squared distribution (White & Liu, 1994). The first experiment exhibits an artificial dataset for which method (b) performs poorly because it is biased towards many-valued attributes, whereas (a) performs well (and so does (c)). The second exhibits another dataset for which method (c) is biased towards towards many-valued attributes and performs poorly (and (b) performs even worse), whereas (a) continues to perform well.

The third subsection presents results for building decision trees on practical datasets using the new method.

3.1 Using the Exact Probability p_f is Biased

In order to show that the exact probability p_f is biased towards attributes with many values, we adopt the experimental setup of White and Liu (1994). This involves an artificial dataset that exhibits no actual association between class and attribute values. For each class, an equal number (300) of instances with random, uniformly distributed attribute values is generated. The estimated p-value of the Freeman and Halton test $\hat{p}$, the exact probability p_f, and the p-value of the parametric chi-squared test p_χ are calculated for this artificial, non-informative, attribute.[5] This procedure is repeated 1000 times with different random seeds used to generate the instances.

Table 1 shows the average values obtained. It can be seen in column (b) that p_f systematically decreases with increasing number of classes and attribute values. Even more importantly, it is always close to zero. If used for pre-pruning at the 0.01 level (as proposed by Martin, 1997), it would fail to stop splitting in every situation except that represented by the first row. On the other hand, neither $\hat{p}$ nor p_χ varies systematically with the number of attribute and class values. For these reasons it is inadvisable to use p_f for attribute selection without preceding it with a significance test.

3.2 Parametric Chi-Squared Test is Biased

A similar experimental procedure was used to show that the parametric chi-squared test is biased in small frequency domains with unevenly distributed samples. Instead of generating the attribute values uniformly, they are skewed so that more samples lie close to the zero point. This is done using the distribution $\lfloor kx^2 \rfloor$, where k is the number of attribute values and x is

[5] Our experiments use $N = 1000$ Monte Carlo trials to estimate $\hat{p}$.

distributed uniformly between 0 and 1. The number of instances is reduced to twenty.

Table 2 shows the average values obtained using this procedure. It can be seen that p_χ decreases systematically as the number of attribute values increases, whereas this is not the case for $\hat{p}$. The test based on p_χ is too liberal in this situation. There also exist situations in which it is too conservative (Good, 1994). If used for pruning in a decision tree, a test that is too liberal does not prune enough, and a test that is too conservative prunes too much.

3.3 Comparison on Practical Datasets

Results are now presented for building decision trees for thirty-one UCI datasets (Merz & Murphy, 1996) using the method described above. We eliminated missing values from the datasets by deleting all attributes with more than 10% missing values, and subsequently removing all instances with missing values. The resulting datasets are summarized in Table 3. All numeric attributes were discretized into four intervals of equal width.[6]

We compare pre-pruned trees built using (a) p_f with prior significance testing using the Freeman and Halton test $\hat{p}$, (b) the exact probability p_f, (c) p_f with prior significance testing using the parametric chi-squared test p_χ, and (d) post-pruned trees built using C4.5's pessimistic pruning with default parameter settings (Quinlan, 1993). We also include results for pruned and unpruned trees as built by C4.5. Note that for (a) and (c) we are now applying the two-step attribute selection procedure developed in Section 2.4, first discarding insignificant attributes and then selecting the best among the remainder. Results are reported for three significance levels: 0.01, 0.05 and 0.10. All results were generated using ten-fold cross-validation repeated ten times with different randomizations of the dataset. The same folds were used for each scheme.[7]

Table 4 shows how method (a) compares with the others. Each row contains the number of datasets for which it builds significantly more (+) or less (−) accurate trees, and significantly smaller (+) or larger (−) trees than the method associated with this row. We speak of results being "significantly different" if the difference is statistically significant at the 1% level according to a paired two-sided t-test, each pair of data points consisting of the estimates obtained in one ten-fold cross-validation run for the two learning schemes being compared. Results are shown for three different significance levels: note that this refers to the level at which attributes are rejected prior to the selection process.

Table 3: Datasets used for the experiments

Dataset	Size	Attributes (numeric/total)	Classes
anneal	898	6/38	5
audiology	216	0/67	24
australian	653	6/15	2
autos	193	14/24	6
balance-scale	625	4/ 4	3
breast-cancer	277	0/ 9	2
breast-w	683	9/ 9	2
german	1000	7/20	2
glass (G2)	163	9/ 9	2
glass	214	9/ 9	6
heart-c	296	6/13	2
heart-h	261	5/10	2
heart-statlog	270	13/13	2
hepatitis	137	3/16	2
hypothyroid	3404	2/24	4
ionosphere	351	34/34	2
iris	150	4/ 4	3
kr-vs-kp	3196	0/36	2
lymphography	148	3/18	4
mushroom	8124	0/21	2
pima-indians	768	8/ 8	2
primary-tumor	336	0/15	21
segment	2310	19/19	7
sick	3404	2/24	2
sonar	208	60/60	2
soybean	630	0/16	15
splice	3190	0/61	3
vehicle	846	18/18	4
vote	312	0/15	2
vowel	990	10/13	11
zoo	101	1/16	7

Observe first that pre-pruning using $\hat{p}$ outperforms pre-pruning using p_f (the three rows marked (b)), confirming our findings from Section 3.1. For all three significance levels $\hat{p}$ dominates p_f in both accuracy and size of the trees produced. These results show that if the splitting attribute is selected based on the value of p_f, it is better to use a significance test first.

One might think that p_f performs poorly with respect to $\hat{p}$ because the former does not prune sufficiently—it is inferior in terms of both accuracy and tree size. Consequently, we also ran pre-pruning using p_f at the 0.005 and 0.001 levels, and found that the performance

[6] If the class information were used when discretizing the attributes, the assumptions of the statistical tests would be invalidated.

[7] Appendix A lists the average accuracy and standard deviation for a representative subset of the methods.

Table 4: Number of times $\hat{p}$ performs significantly better (+) or worse (−) than (b) p_f, (c) p_χ, (d) post-pruned trees, and pruned and unpruned C4.5 trees with respect to accuracy and tree size

		Accuracy		Tree Size	
	$\hat{p}$	+	−	+	−
$p_{\text{fixed}} = 0.01$	(b) p_f	8	5	17	6
	(c) p_χ	9	3	8	11
	(d) post-pruned	4	14	20	7
	C4.5 pruned	3	17	20	7
	C4.5 unpruned	11	11	31	0
$p_{\text{fixed}} = 0.05$	(b) p_f	8	2	22	3
	(c) p_χ	6	6	24	2
	(d) post-pruned	4	9	8	17
	C4.5 pruned	2	16	11	15
	C4.5 unpruned	8	9	29	2
$p_{\text{fixed}} = 0.1$	(b) p_f	9	2	24	1
	(c) p_χ	5	5	24	0
	(d) post-pruned	4	12	5	22
	C4.5 pruned	3	16	3	24
	C4.5 unpruned	8	8	29	2

Table 5: Number of times $\hat{p}$ with gain ratio (Method a′) performs significantly better (+) or worse (−) than $\hat{p}$ with p_f (Method a), and pruned and unpruned C4.5 trees

		Accuracy		Tree Size	
	$\hat{p}$ with gain ratio	+	−	+	−
$p_{\text{fixed}} = 0.01$	$\hat{p}$ with p_f	8	3	10	10
	C4.5 pruned	3	14	21	6
	C4.5 unpruned	13	7	30	1
$p_{\text{fixed}} = 0.05$	$\hat{p}$ with p_f	10	4	11	14
	C4.5 pruned	0	10	10	14
	C4.5 unpruned	12	7	30	1
$p_{\text{fixed}} = 0.1$	$\hat{p}$ with p_f	10	5	11	12
	C4.5 pruned	1	15	6	22
	C4.5 unpruned	13	8	30	0

difference between p_f and $\hat{p}$ can not be eliminated by adjusting the significance level.

Next, observe from the three rows marked (c) that for the 0.01 significance level, pre-pruning using $\hat{p}$ beats pre-pruning using p_χ with respect to the accuracy of the resulting trees. For this significance level the two methods produce trees of similar size. However, for both the 0.05 and the 0.1 levels $\hat{p}$ produces trees that are significantly smaller than those produced by p_χ. For these two significance levels the two methods perform comparably as far as accuracy is concerned. These facts indicate that for both the 0.05 and the 0.1 levels p_χ is a more liberal test than $\hat{p}$ if applied to attribute selection and pre-pruning; p_χ stops later than $\hat{p}$—as for the artificial dataset used in Section 3.2. However, it is sometimes more conservative—in particular for the 0.01 level. The two tests really do behave differently: they cannot be forced to behave in the same way by adjusting their significance levels. However, the results show that trees produced by $\hat{p}$ are preferable to those produced by p_χ.

Table 4 also shows that post-pruning consistently beats pre-pruning using $\hat{p}$, so far as accuracy is concerned (rows marked (d)). Our findings show that all the investigated pre-pruning methods perform significantly worse than pessimistic post-pruning.[8] For both the 0.01 and the 0.05 levels, there are five datasets

[8] This contradicts a previous result (Martin, 1997) that trees pre-pruned using p_f are as accurate as, and smaller than, trees post-pruned using pessimistic pruning.

on which *all* pre-pruning methods consistently perform significantly worse than post-pruning: hypothyroid, kr-vs-kp, sick, splice, and vowel. On kr-vs-kp and vowel the pre-pruning methods stop too early, on the other three they stop too late. This means that the problem cannot be solved by adjusting the significance level of the pre-pruning methods.

For reference Table 4 also includes results for pruned and unpruned decision trees built by C4.5. C4.5's method for building pruned trees differs from post-pruning method (d) only in that it employs the gain ratio[9] instead of p_f for attribute selection.

Suprisingly, Table 4 shows that $\hat{p}$ does not perform better than C4.5's unpruned trees as far as accuracy is concerned, although $\hat{p}$ performs better than unpruned trees built using p_f (results not shown). This indicates that the gain ratio produces more accurate trees than p_f. We therefore replaced attribute selection using p_f in the second step of pre-pruning method (a) by selection based on the gain ratio. As Table 5 shows, the new method (a′)—selection based on the gain ratio with prior significance testing using the Freeman and Halton test $\hat{p}$—indeed performs better than method (a), and it also outperforms C4.5's unpruned trees. However, as Table 5 also shows, post-pruning—in this case represented by C4.5's pruned trees—still consistently beats pre-pruning using $\hat{p}$.

4 Related Work

Several researchers have applied parametric statistical tests to attribute selection in decision trees (White & Liu, 1994; Kononenko, 1995) and proposed remedies for their shortcomings (Martin, 1997). These are reviewed in the next section. Following that we discuss work on permutation tests for machine learning, none of which has been concerned with attribute selection in decision trees.

4.1 Use of Statistical Tests for Attribute Selection

White and Liu (1994) compare several entropy-based selection criteria to parametric tests that rely on the chi-squared distribution. More specifically, they compared the entropy-based measures to parametric tests based on both the chi-squared and log likelihood ratio statistics. They conclude that each of the entropy measures favors attributes with larger numbers of values, whereas the statistical tests do not suffer from this problem. However, they also mention the problem of small expected frequencies with parametric tests and suggest the use of Fisher's exact test as a remedy. The extension of Fisher's exact test to $r \times c$ tables is the Freeman and Halton test that we have used above.

Kononenko (1995) repeated and extended these experiments and investigated several other attribute selection criteria as well. He shows that the parametric test based on the log likelihood ratio is biased towards attributes with many values if the number of classes and attribute values relative to the number of instances exceed the corresponding figures considered by White and Liu (1994). This is not surprising: it can be traced to the problem of small expected frequencies. For the log likelihood ratio the effect is more pronounced than for the chi-squared statistic (Agresti, 1990).

Kononenko also observes another problem with statistical tests. The restricted floating-point precision of most computer arithmetic makes it difficult to use them to discriminate between different *informative* attributes. The reason for this is that the association to the class is necessarily highly significant for all informative attributes.[10] However, there is an obvious solution, which we pursue in this paper: once it has been established that an attribute is significant, it can be compared to other significant attributes using an attribute selection criterion that measures the strength of the association.

Recently, Martin (1997) used the exact probability of a contingency table given its marginal totals p_f for attribute selection and pre-pruning. Our method differs from his only in that we employ a significance test, based on p_f but not identical to it, to determine the significance of an attribute before selecting the best of the *significant* attributes according to p_f. As Section 3 of this paper establishes, direct use of p_f for attribute selection produces biased results.

4.2 Use of Permutation Tests in Machine Learning

Apparently the first to use a permutation test for machine learning, Gaines (1989) employs an approximation to Fisher's exact test to judge the quality of rules found by the INDUCT rule learner.[11] Instead of the

[9] More precisely, it selects the attribute with maximum gain ratio among the attributes with more than average information gain.

[10] The probability that the null hypothesis of no association between attribute and class values is incorrectly rejected is very close to zero.

[11] He uses the one-tailed version of Fisher's exact test.

Figure 1: Two 2×2-tables which both optimize the test statistic

3	0
0	3

0	3
3	0

hypergeometric distribution he uses the binomial distribution, which is a good approximation if the sample size is small relative to the population size (smaller than 10 percent).

Jensen (1992) gives an excellent introduction to permutation tests.[12] He discusses several alternatives, points out their weaknesses, and deploys the methodology in a prototypical rule learner. However, he does not mention the prime advantage of permutation tests, which makes them especially interesting in the context of decision trees: their applicability to small-frequency domains.

5 Conclusions

We have applied an approximate permutation test based on the multiple hypergeometric distribution to attribute selection and pre-pruning in decision trees, and explained why it is preferable to tests based on the chi-squared distribution. We have shown that using the exact probability of a contingency table given its marginal totals without a prior significance test is biased towards attributes with many values and performs worse in comparison. Although we were able to improve on existing methods for pre-pruning, we could not achieve the same accuracy as post-pruning.

Apart of the standard explanation that pre-pruning misses hidden attribute interactions, there are two other possible reasons for this result. The first is that we did not adjust for multiple comparisons when testing the significance of an attribute. Recently, Jensen and Schmill (1997) showed how to reduce the size of a post-pruned tree significantly by taking multiple hypotheses into account using a technique known as the "Bonferroni correction." The second reason is that tests for $r \times c$ contingency tables are inherently multi-sided. Consider the table shown at the left of Figure 1, which corresponds to a perfect classification of two classes using an attribute with two values. There is another permutation of class labels, shown at the right, that also results in a contingency table with the same optimum value of the test statistic. The significance level achieved by the original table is only half as great as it would be if there were only one table that optimized the test statistic. In the case of two attributes and two classes, the one-sided version of Fisher's exact test avoids this problem. Generalizing this to the $r \times c$ case appears to be an open problem.

[12] He uses the term "randomization test" instead of permutation test.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley & Sons.

Gaines, B. (1989). An ounce of knowledge is worth a ton of data. In *Proceedings of the 6th International Workshop on Machine Learning* (pp. 156–159). Morgan Kaufmann.

Good, P. (1994). *Permutation Tests.* New York: Springer-Verlag.

Jensen, D. (1992). *Induction with Randomization Testing.* PhD thesis, Washington University, St. Louis, Missouri. [http://eksl-www.cs.umass.edu/~jensen/papers/dissertation.ps].

Jensen, D. & Schmill, M. (1997). Adjusting for multiple comparisons in decision tree pruning. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining.* AAAI Press. [http://eksl-www.cs.umass.edu/~jensen/papers/kdd97.ps].

Kononenko, I. (1995). On biases in estimating multi-valued attributes. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1034–1040). Morgan Kaufmann.

Martin, J. K. (1997). An exact probability metric for decision tree splitting and stopping. *Machine Learning, 28(2,3)*, 257–291.

Merz, C. J. & Murphy, P. M. (1996). *UCI Repository of Machine Learning Data-Bases.* Irvine, CA: University of California, Department of Information and Computer Science. [http://www.ics.uci.edu/~mlearn/MLRepository.html].

Press, W. H., Teukolsky, S., Vettering, W. & Flannery, B. (1988). *Numerical Recipes in C* (2nd Ed.). Cambridge University Press.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

White, A. P. & Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Machine Learning, 15(3)*, 321–329.

A Accuracy for Practical Datasets

Table 6: Experimental results: percentage of correct classifications, and standard deviation using $\hat{p}$, p_f, p_χ, post-pruned trees, $\hat{p}$ with gain ratio, C4.5's pruned trees, and C4.5's unpruned trees. Because of space constraints, we could only include results for one of the three p_{fixed} vales used in Table 4: we chose $p_{\text{fixed}} = 0.05$. In the last six columns, figures are marked with • if they are significantly worse than the corresponding results for $\hat{p}$, and with ◦ if they are significantly better.

	$\hat{p}$	p_f	p_χ	post-pruned	$\hat{p}$ with gain ratio	C4.5 pruned	C4.5 unpruned
anneal	98.6±0.1	98.5±0.0 •	99.0±0.1 ◦	98.4±0.1 •	98.3±0.3	98.0±0.3 •	98.3±0.3
audiology	71.6±1.9	70.3±1.9 •	71.5±1.7	71.9±1.3	73.8±1.2 ◦	74.8±1.0 ◦	74.8±1.3 ◦
australian	85.7±0.5	86.7±0.5 ◦	85.0±0.5 •	86.4±0.0 ◦	84.8±0.5 •	85.2±0.4	83.8±1.0 •
autos	67.3±2.2	67.2±2.4	72.7±2.4 ◦	70.5±2.4 ◦	73.3±2.3 ◦	73.0±2.0 ◦	72.9±2.3 ◦
balance-scale	66.1±0.9	70.5±1.2 ◦	65.9±1.2	67.3±1.0	67.2±1.2 ◦	67.9±1.0 ◦	74.1±1.0 ◦
breast-cancer	69.0±1.5	65.0±1.4 •	69.8±1.2	67.6±1.1	72.5±1.1 ◦	74.4±1.2 ◦	66.6±1.4 •
breast-w	95.2±0.7	95.1±0.6	95.0±0.7	95.2±0.6	95.7±0.3	96.0±0.3 ◦	95.6±0.3
german	70.3±0.7	70.4±0.7	70.4±1.1	70.5±0.5	70.5±0.8	70.9±0.8	67.2±1.2 •
glass (G2)	70.5±4.3	70.6±2.5	70.5±3.3	71.3±1.7	67.3±2.5	79.7±1.4 ◦	79.5±1.6 ◦
glass	59.8±1.4	59.3±1.4	59.6±1.1	60.2±1.3	60.1±1.6	59.9±2.1	59.3±1.4
heart-c	78.2±1.1	76.8±1.4	76.6±0.9 •	79.2±2.4	77.0±1.2	77.5±1.2	75.1±1.4 •
heart-h	73.9±0.9	72.6±1.6	74.8±1.2	73.7±0.9	77.8±1.2 ◦	79.5±0.8 ◦	76.6±1.0 ◦
heart-statlog	79.2±1.5	77.7±1.7 •	78.1±1.9 •	80.1±0.7	76.2±1.6 •	78.5±1.9	75.7±2.0 •
hepatitis	79.8±2.4	79.5±2.2	79.5±1.7	80.7±1.6	84.4±1.8 ◦	84.4±1.3 ◦	80.7±1.4
hypothyroid	91.7±0.1	91.7±0.0	91.7±0.0	91.9±0.0 ◦	91.7±0.0	91.9±0.0 ◦	91.7±0.1
ionosphere	87.0±1.0	86.7±0.8	87.4±0.8	88.1±0.5 ◦	87.8±1.4	87.2±0.6	86.6±0.7
iris	91.8±0.3	91.5±0.9	91.8±0.3	91.5±0.8	91.9±0.2	91.5±0.9	90.7±1.1
kr-vs-kp	99.3±0.1	99.3±0.1	99.3±0.1	99.4±0.1 ◦	99.3±0.1	99.5±0.1 ◦	99.5±0.1 ◦
lymphography	75.2±0.8	76.3±2.1	75.2±1.5	76.0±2.4	76.1±1.6	78.6±1.6 ◦	75.8±2.0
mushroom	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
pima-indians	74.0±0.8	72.9±0.7	74.2±0.5	71.9±0.4 •	74.1±0.6	74.1±0.5	69.4±0.8 •
primary-tumor	39.8±1.1	36.1±1.4 •	37.6±1.4 •	35.7±1.4 •	38.7±1.9	40.0±0.5	40.3±1.1
segment	91.0±0.2	91.2±0.3	91.1±0.2 ◦	91.3±0.2	91.5±0.3 ◦	91.8±0.2 ◦	91.8±0.3 ◦
sick	93.3±0.1	93.3±0.1	93.2±0.1 •	93.4±0.0 ◦	93.3±0.1	93.4±0.0 ◦	93.2±0.1 •
sonar	68.8±2.5	68.3±2.5	68.6±3.5	69.1±2.4	70.3±2.6	71.5±2.2	70.5±3.1
soybean	75.1±0.8	72.2±0.8 •	76.1±0.7 ◦	73.5±0.6 •	77.6±0.5 ◦	77.7±0.5 ◦	76.7±0.7 ◦
splice	92.6±0.3	92.3±0.3 •	92.2±0.3 •	93.4±0.2 ◦	93.2±0.2 ◦	94.2±0.2 ◦	92.2±0.2 •
vehicle	63.4±0.9	62.0±0.6 •	64.1±1.0 ◦	64.2±0.7	65.7±0.7 ◦	66.1±0.5 ◦	64.2±0.7
vote	95.4±0.4	95.5±0.4	95.5±0.3	95.6±0.5	95.5±0.4	95.5±0.4	96.2±0.5 ◦
vowel	77.9±1.0	78.0±1.0	79.5±1.0 ◦	80.8±1.0 ◦	73.8±0.6 •	76.6±0.5 •	78.2±0.7
zoo	92.5±1.8	92.8±1.6	94.0±2.0	94.8±2.1 ◦	89.6±1.4 •	90.8±1.5	91.5±1.4