Machine Learning, 8, 75–85 (1992)
© 1992 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

# *Technical Note*
# A Further Comparison of Splitting Rules for Decision-Tree Induction

WRAY BUNTINE[1]
TIM NIBLETT (TIM@TURING.AC.UK)
*The Turing Institute, George House, 36 North Hanover St., Glasgow, G1 2AD, U.K.*

**Abstract.** One approach to learning classification rules from examples is to build decision trees. A review and comparison paper by Mingers (Mingers, 1989) looked at the first stage of tree building, which uses a "splitting rule" to grow trees with a greedy recursive partitioning algorithm. That paper considered a number of different measures and experimentally examined their behavior on four domains. The main conclusion was that a random splitting rule does not significantly decrease classificational accuracy. This note suggests an alternative experimental method and presents additional results on further domains. Our results indicate that random splitting leads to increased error. These results are at variance with those presented by Mingers.

**Keywords.** Decision trees, induction, noisy data, comparative studies

## 1. Introduction

Empirical comparisons are an important part of machine learning research (Kibler & Langley, 1988). In the area of decision tree induction empirical comparisons have been widely used by Breiman, et al. (Breiman, et al., 1984), Quinlan and others (Quinlan, 1988; Quinlan, 1989; Cestnik, et al., 1987) to guide the development of these learning systems. Several comparisons of algorithms from different learning paradigms have been presented recently (Mooney, et al., 1989; Weiss & Kapouleas, 1989; Fisher & McKusick, 1989). These studies and the debate that followed their presentation have highlighted just how difficult it is to be thorough and fair when comparing algorithms.

This note presents an empirical comparison of a specialized aspect of decision tree induction. Trees are induced in a two-stage process often referred to as *growing* and *pruning*. Growing involves using training examples to build a tree. This stage often over-grows, in the sense that the induced trees are too large. Such trees are said to "track noise" in the data. The second stage prunes back the tree to a smaller and usually more accurate classifier of future examples. The growing process uses a procedure referred to as recursive partitioning to grow a tree with a one-ply lookahead to select the best test at each node. This "greedy" algorithm fixes the current "best" test at the current node and goes on to grow subtrees from that node, without subsequent backtracking. A *splitting rule* is a one-ply lookahead heuristic used to guess the "best" test to make at the current node in the tree.

[1]Current address: Research Institute for Advanced Computer Science and Artificial Intelligence Research Branch, NASA Ames Research Center, Mail Stop 269-2, Moffett Field, CA 94035, USA. Email: wray@ptolemy.arc.nasa.gov.

A recent paper by Mingers (Mingers, 1989) undertook an empirical comparison of different splitting rules. We perceived a potential problem with the comparisons reported by Mingers and, noting Mingers' concern that the experimental work be confirmed and extended on different domains, have undertaken a further comparison in the spirit of confirming and extending earlier experimental results.

In what follows we assume familiarity with Mingers' paper (Mingers, 1989). Mingers provides an introduction to the splitting rules under evaluation, then describes his experimental procedure and the results obtained. His main conclusion is that the predictive error of induced trees is not sensitive to the splitting rule, and in particular that use of a random splitting rule does not significantly increase classification error (Mingers, 1989, p. 338).

We begin in Section 2 by discussing Mingers' experimental method and more delicate aspects of the comparison of learning techniques, to arrive at our own experimental method. Then we present results from our own experiments with additional data sets in Sections 3 to 5. These results indicate that a random splitting rule does perform significantly worse than other methods. Section 6 presents our conclusions.

## 2. Experimental method

Mingers' experimental method is summarized in the following quotation (Mingers, 1989, p. 334)

> To obtain independent test data and reliable results, each original data set was split randomly (70/30) into a training and a test set. The trees were grown and pruned on the training set and then error was measured on the test set. In fact, the test set was not wholly independent since it is used in Breiman's pruning method. This develops a number of pruned trees entirely from the training set, but then selects the best via the test set.

Mingers used four data sets in his comparison and checked significance by applying ANOVA (two way analysis of variance) tests to the matrix of error averages obtained. We have several concerns with this experimental method. In the interests of promoting discussion on comparative studies, we discuss them at length here.

1. The error estimates were obtained from the same test set used for pruning. Since the pruning method essentially finds the pruned subtree giving minimum errors, the error estimates will be downwardly biased, under-estimating error. Although we are only interested in comparative error, some trees may give more scope for pruning than others, and thus the comparative estimates may be biased. Mingers acknowledged this problem (see quote above), but did not report whether tests were made to check if it caused detrimental effects.
2. The comparison was made on only four test sets. While this is adequate for evaluating an algorithm with a strong theoretical backing, or an algorithm that operates on a limited range of problems, we felt it was not adequate to evaluate eleven heuristics such as the splitting rules. Our belief is supported by the non-significance of the ANOVA test ran by Mingers. While the splitting rules are based on seemingly solid statistical tests, all

the tests are "correct" only under assumptions quite different from those that apply when growing trees. As such, their justification is heuristic at best.

3. The use of ANOVA significance testing, while a useful way of checking overall significance of results presented as tables, is not the most appropriate for the task here. The ANOVA test makes the explicit assumption that standard deviations for the different quantities reported, in this case error averages, are constant. This is certainly not the case here where standard deviations for the error averages vary widely. From our own estimates these range from as little as 0.05% for the hypo data to as much as 2.6% for some of the other domains. This also means, of course, that the total of the error averages across the domains (the right hand column in Mingers' Table 10) may be a poor indicator of relative merit of the splitting rules. While ANOVA testing might be adequate for a rough check in this case, we felt it should be complemented with another significance test.

   We chose the paired $t$-test to compare differences between individual pairs of error averages. Although this does not provide a global significance test on relative merit of the splitting rules, it is appropriate for individual comparisons.

For our experiments, we chose for reasons of expediency to evaluate only four splitting rules. These were the GINI index of diversity (Breiman, et al., 1984), information gain (Quinlan, 1986), the Marshall correction (Mingers, 1989) and a random selection of attribute for splitting. The random selection randomly selected a cut-point if a real-valued attribute was being tested.

A variety of data sets, covering a broad range of domain characteristics, were chosen from public-domain[1] data bases. We were hoping to get data sets with different mixtures of attribute types (real, binary, multi-valued discrete, etc.), with different numbers of classes, with different proportions of classes, some with rare classes and some with equiprobable classes, and applications of different styles such as medical data, artificial data, diagnosis data, control data, etc.

Our modified experimental method, applied to each data set, is described below. We have also included details of the statistical tests made.

1. Repeat twenty times:
   (a) The original data set is split randomly into two parts, a training set and a test set, where the training set is always a fixed size (see Table 1). The training set is further split into two parts (70%/30%) referred to as the growing set and pruning set.
   (b) For each splitting rule:
      i. Grow a tree using the growing set and the splitting rule under evaluation.
      ii. Prune the tree using Breiman, et al.'s "cost complexity pruning with test set and the 0-SE rule" (Breiman, et al., 1984) using the pruning set.
      iii. Estimate the error using the test set.
2. Determine the sample average and standard deviation for the error estimates for each splitting rule.
3. Determine the $t$-value for the paired $t$-test between different pairs of splitting rules.

For pruning, we chose to use Breiman, et al.'s "cost complexity pruning with test set" (Breiman, et al., 1984) using the 0-SE rule. Breiman, et al. claim this should give more

accurate but larger trees than their 1-SE rule. We verified this by reproducing the experiments with the 1-SE rule, as discussed later. The 70%/30% split used for growing and pruning in Step 1(a) was chosen, because it is often used in the literature. We made no attempt at changes.

## 3. Data sets

The data sets used are described below. Some versions of these are also used by Mingers.

**hypo** The hypothyroid data set is Quinlan's hypothyroid data described in (Quinlan, et al., 1987; Quinlan, 1988), and supplied from an expert system for advising on thyroid disorders that is in routine use at the Garvan Institute of Medical Research in Sydney. The data set of 3772 examples records expert opinion on possible hypothyroid conditions from 29 real and discrete attributes of the patient such as sex, age, taking of relevant drugs, and hormone readings taken from drug samples. Unknowns exist in the data. This is a fairly simple classification task, as the expert opinions are reliable, "noise" is often expert mistakes, and a major part of predicting the hypothyroid condition comes from the level of thyroid hormone in the blood sample (one attribute).

**breast** The "breast" data set comes from the breast cancer domain in oncology. The classes are reoccurrence or non-reoccurrence of breast cancer sometime after an operation. There are nine attributes giving details about the original cancer nodes, position on the breast, and age, with multi-valued discrete and real values. This data set, along with the next two, comes from the Institute of Oncology, Ljubljana and has been previously reported on by Cestnik, et al. (Cestnik, et al., 1987) and Clark and Niblett (Clark & Niblett, 1989).

**tumor** The "tumor" data set gives examples of the location of a primary tumor. There are twenty-two class values, a few unknown attribute values in the most important attributes, several multi-valued discrete attributes, but mainly binary attributes.

**lymph** The "lymph" data set comes from the lymphography domain in oncology. The classes are normal, metastases, malignant, and fibrosis, and there are nineteen attributes giving details about the lymphatics and lymph nodes, with multi-valued discrete and real values but no unknowns.

**LED** The "LED" data set is Breiman, et al.'s classic manufactured test data on the digit recognition problem (Breiman, et al., 1984). There are ten classes, representing whether a faulty LED is showing 0–9. The seven binary valued attributes record whether each of the seven LED elements (one on top and on bottom, two on each side and one horizontally in the center) is on or off. The LED is made faulty by adding 10% noise independently to each element. The corresponding prediction task has a theoretical minimum error of about 27.3%, with approximately one half the digits being correctly represented. The corresponding learning task is actually best solved using a simple Bayes classifier (which assumes attributes are independent given class (Buntine, 1989)). All samples of this task were drawn from a population of 3000.

**mush** The "mush" data set records whether mushrooms from the Agaricus and Lepiota families are poisonous or edible, given details about the mushrooms in twenty-two discrete attributes describing cap shape and surface, etc. The data was transcribed from a field manual on mushrooms and filled in to give 8124 examples by J. Schlimmer and reported in his thesis and elsewhere (Schlimmer & Granger, Jr., 1986). Due to its book source, it can be considered virtually noise free.

**votes** The "votes" data extracts details from the 1984 United States congressional voting records. These 435 examples record key votes of 267 democrats and 168 republicans on issues such as adoption of the budget, immigration, and education spending. Votes have been simplified to yea, nay or unknown (this is a third discrete value and is not treated as an unknown or undetermined attribute value). This data set was also originally transcribed by J. Schlimmer.

**votes1** This data set was derived from the previous data set by deleting the most significant attribute *physician-fee-freeze*. This follows a suggestion by Donald Michie. The most accurate and reliable tree for the full "votes" data set is the trivial tree of depth one which has a single test on *physician-fee-freeze*.

**iris** The "iris" data set, Fisher's classic test data (Fisher, 1936), describes three classes of iris plants using the real valued attributes petal and sepal width and petal and sepal length, with no unknowns. This data set of 150 examples gives good results with almost all classical learning methods (discriminant analysis, etc.) and is included here for comparison.

**glass** This data set represents the problem of identifying glass samples taken from the scene of an accident. The 214 examples were originally collected by B. German of the Home Office Forensic Science Service at Aldermaston, Reading, Berkshire in the UK. There are seven classes such as building or vehicle windows, containers, tableware, and headlamps. The nine attributes are all real valued and fully known, representing refractive index and the percent weight of oxides such as silicon and aluminum.

**xd6** The "xd6" data set, of 600 examples, has ten boolean attributes $A1$–$A10$ with class given by the boolean formula

$$A1 \wedge A2 \wedge A3 \vee A4 \wedge A5 \wedge A6 \vee A7 \wedge A8 \wedge A9.$$

The class also has 10% class noise added, so the optimal prediction error is 10%.

**pole** The pole data set records the experience of a human on a simple one-dimensional pole balancing task. This was done with a moderately experienced human using a graphical simulation on a PC. The data has been recorded as part of a much larger comparative (human and machine) study of learning by the Comparative Studies Group at the Turing Institute (Michael Bain, Donald Michie and Jean Hayes-Michie). The pole is balanced using simple bang-bang control. The data set records 4 real values (current angle, x-position, and their rates of change). The class is whether the human subject made a left or right bank. Over 1800 different circumstances are recorded.

*Table 1.* Properties of the data sets.

| Data Set | Classes | Attr.s | Real | Multi | % Unkn | Training Set | Test Set | % Base Error |
|---|---|---|---|---|---|---|---|---|
| hypo | 4 | 29 | 7 | 1 | 5.5 | 1000 | 2772 | 7.7 |
| breast | 2 | 9 | 4 | 2 | 0.4 | 200 | 86 | 29.7 |
| tumor | 22 | 18 | 0 | 3 | 3.7 | 237 | 102 | 75.2 |
| lymph | 4 | 18 | 1 | 8 | 0 | 103 | 45 | 45.3 |
| LED | 10 | 7 | 0 | 0 | 0 | 200 | 1800 | 90.0 |
| mush | 2 | 22 | 0 | 18 | 0 | 200 | 7924 | 48.2 |
| votes | 2 | 17 | 0 | 17 | 0 | 200 | 235 | 38.6 |
| votes1 | 2 | 16 | 0 | 16 | 0 | 200 | 235 | 38.6 |
| iris | 3 | 4 | 4 | 0 | 0 | 100 | 50 | 66.7 |
| glass | 7 | 9 | 9 | 0 | 0 | 100 | 114 | 64.5 |
| xd6 | 2 | 10 | 0 | 0 | 0 | 200 | 400 | 35.5 |
| pole | 2 | 4 | 4 | 0 | 0 | 200 | 1647 | 49.0 |

Some data sets were obtained through indirect sources. The "breast," "tumor" and "lymph" data sets were originally collected at the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia, in particular by G. Klajnšček and M. Soklic (lymphography data), and M. Zwitter (breast cancer and primary tumor). The data was converted into easy-to-use experimental material by Igor Kononenko, Faculty of Electrical Engineering, Ljubljana University. The data has been the subject of a series of comparative studies, for instance (Cestnik, et al., 1987). The hypothyroid data ("hypo") came originally from the Garvan Institute of Medical Research, Sydney. The data sets "glass," "votes" and "mush" came from David Aha's Machine Learning Database available over the academic computer network from the University of California at Irvine, "hypo" and "xd6" came from a collection by Ross Quinlan of the University of Sydney (Quinlan, 1988), "breast," "lymph" and "tumor" came via Pete Clark of the Turing Institute, and "iris" from Stuart Crawford of Advanced Decision Systems. Versions[2] of the last four mentioned data sets are also available from the Irvine Machine Learning Database.

Major properties of the data sets are given in Table 1. Columns headed "real" and "multi" are the number of attributes that are treated as real-valued or ordered and as multi-valued discrete attributes respectively. Percentage unknown is the proportion of all attribute values that are unknown. These are usually concentrated in a few attributes. Percentage base error is the percentage error obtained if the most frequent class is always predicted. Good trees should give a significant improvement over this.

## 4. Implementation

The decision tree implementation used in these experiments was originally written by David Harper, Chris Carter, and other students at the University of Sydney from 1984 to 1988. The present version has been largely rewritten by Wray Buntine. Performance of the current system was compared to earlier versions to check that bugs were not introduced during rewriting. Unknown attribute values were treated as follows. When evaluating a test, an example with unknown outcome had its unit weight split across outcomes according to

the proportion found for examples of the same class. When partitioning examples, an example with unknown outcome was passed down the most frequent branch. When classifying a new example, an example with unknown outcome was passed down each branch with weight proportional to the number of examples in the training set passed down the branch.

## 5. Results

Leaf counts and average errors for pruned trees grown as described above are given in Tables 2 and 3 respectively.

These results are given in the form "29.7 ± 3.4." This first figure means that the average on the test set (the full data set minus the training set) for the 20 trials was 29.7%. The

*Table 2.* Leaf count of pruned trees for different splitting rules.

| | Splitting Rule | | | |
|---|---|---|---|---|
| Data Set | GINI | Info. Gain | Marsh. | Random |
| hypo | 5.0 ± 1.2 | 4.8 ± 1.3 | 5.8 ± 1.3 | 34.0 ± 14.6 |
| breast | 10.2 ± 7.1 | 9.3 ± 6.8 | 6.0 ± 4.1 | 25.4 ± 10.0 |
| tumor | 19.6 ± 5.8 | 22.5 ± 5.4 | 17.7 ± 6.2 | 32.8 ± 11.4 |
| lymph | 8.2 ± 5.0 | 7.5 ± 3.8 | 7.7 ± 3.2 | 15.5 ± 8.0 |
| LED | 13.3 ± 2.7 | 13.0 ± 1.9 | 13.1 ± 1.7 | 19.4 ± 4.7 |
| mush | 12.4 ± 5.2 | 12.4 ± 5.2 | 23.3 ± 8.1 | 48.7 ± 21.5 |
| votes | 5.1 ± 2.5 | 5.2 ± 2.6 | 12.4 ± 6.0 | 15.9 ± 8.9 |
| votes1 | 8.9 ± 4.0 | 9.4 ± 5.6 | 13.0 ± 5.5 | 22.9 ± 10.2 |
| iris | 3.5 ± 0.5 | 3.5 ± 0.5 | 3.4 ± 0.7 | 12.1 ± 5.7 |
| glass | 8.1 ± 2.4 | 8.9 ± 1.8 | 8.5 ± 2.8 | 21.8 ± 6.5 |
| xd6 | 14.9 ± 3.6 | 14.8 ± 3.8 | 14.8 ± 3.9 | 20.1 ± 5.1 |
| pole | 5.7 ± 4.0 | 5.8 ± 3.4 | 5.4 ± 2.9 | 22.7 ± 8.2 |

*Table 3.* Error for different splitting rules (pruned trees).

| | Splitting Rule | | | |
|---|---|---|---|---|
| Data Set | GINI | Info. Gain | Marsh. | Random |
| hypo | 1.01 ± 0.29 | 0.95 ± 0.22 | 1.27 ± 0.47 | 7.44 ± 0.53 |
| breast | 28.66 ± 3.87 | 28.49 ± 4.28 | 27.15 ± 4.22 | 29.65 ± 4.97 |
| tumor | 60.88 ± 5.44 | 62.70 ± 3.89 | 61.62 ± 3.98 | 67.94 ± 5.68 |
| lymph | 24.44 ± 6.92 | 24.00 ± 6.87 | 24.33 ± 5.51 | 32.33 ± 11.25 |
| LED | 33.77 ± 3.06 | 32.89 ± 2.59 | 33.15 ± 4.02 | 38.18 ± 4.57 |
| mush | 1.44 ± 0.47 | 1.44 ± 0.47 | 7.31 ± 2.25 | 8.77 ± 4.65 |
| votes | 4.47 ± 0.95 | 4.57 ± 0.87 | 11.77 ± 3.95 | 12.40 ± 4.56 |
| votes1 | 12.79 ± 1.48 | 13.04 ± 1.65 | 15.13 ± 2.89 | 15.62 ± 2.73 |
| iris | 5.00 ± 3.08 | 4.90 ± 3.08 | 5.50 ± 2.59 | 14.20 ± 6.77 |
| glass | 39.56 ± 6.20 | 50.57 ± 6.73 | 40.53 ± 6.41 | 53.20 ± 5.01 |
| xd6 | 22.14 ± 3.23 | 22.17 ± 3.36 | 22.06 ± 3.37 | 31.86 ± 3.62 |
| pole | 15.43 ± 1.51 | 15.47 ± 0.88 | 15.01 ± 1.15 | 26.38 ± 6.92 |

*Table 4.* Difference and significance of error for GINI splitting rule versus others.

| | Splitting Rule | | |
|---|---|---|---|
| Data Set | Info. Gain | Marsh. | Random |
| hypo | −0.06 (0.82) | 0.26 (0.99) | 6.43 (1.00) |
| breast | −0.17 (0.23) | −1.51 (0.94) | 0.99 (0.72) |
| tumor | 1.81 (0.84) | 0.74 (0.39) | 7.06 (0.99) |
| lymph | −0.44 (0.83) | −0.11 (0.05) | 7.89 (0.99) |
| LED | 0.12 (0.17) | 0.38 (0.41) | 5.41 (0.99) |
| mush | 0.00 (0.00) | 5.86 (1.00) | 7.32 (0.99) |
| votes | 0.11 (0.55) | 7.30 (0.99) | 7.94 (0.99) |
| votes1 | 0.26 (0.47) | 2.34 (0.98) | 2.83 (0.99) |
| iris | −0.10 (0.67) | 0.50 (0.90) | 9.20 (0.99) |
| glass | 1.01 (0.50) | 0.96 (0.53) | 13.64 (0.99) |
| xd6 | 0.04 (0.11) | −0.07 (0.20) | 9.72 (0.99) |
| pole | 0.03 (0.11) | −0.43 (0.83) | 10.95 (0.99) |

second figure means that the sample standard deviation of this figure is 3.4%. This gives an idea of how much the quantity varied from sample to sample. The sample standard deviation for error also contains a residual element due to the fact that error is an estimation from a sometimes small test set. Bear in mind this residual element is constant across tree growing methods because training/test data sets are identical for each method.

Significance testing using the two-tailed paired $t$-test is reported in Table 4.

All significance results are given in a form such as 0.53 (0.21). The first number is the *average difference in errors* between the second and first methods, calculated as

$$\frac{1}{|trials|} \sum_{p \in trials} (error\text{-}2_p - error\text{-}1_p).$$

where $error\text{-}1_p$ is the error for the $p$-th trial for the 1-st method, etc. Bear in mind there were 20 trials. The second number is the significance of this difference according to the two-tailed paired $t$-test. This is done by first constructing a t-value on whether the average of the random variable $error\text{-}2_p - error\text{-}1_p$ differs from 0, and then determining the significance of this value according to the two-tailed $t$-test. For instance, a result of the form 0.53 (0.99) means the average error is less for GINI splitting with significance of greater than 99%, a result of the form −0.53 (0.86) means the average error is greater for GINI splitting with significance of greater than 86%, and a result with difference of 0.00 always has a significance of 0%, because we have no evidence that it is greater or less. Sometimes a significance of 100% is reported. In these cases, the $t$ value was so large that the significance level is more than 99.9%.

If we require a significance level of 90%, then the random splitting rule is inferior to GINI in 11 of the 12 domains, the Marshall correction is inferior to GINI in 4 domains and superior in 1 domain out of the 12, and the information gain criteria is statistically indistinguishable from the GINI criteria.

We also produced similar tables for two other pruning methods by Breiman, et al.: cost complexity with test set and 1-SE rule and cost complexity with 10-fold cross validation and 1-SE rule. As suggested by Mingers, these did not change our overall conclusions significantly, although individual error averages changed quite widely for the different pruning methods. For instance, on the digit LED domain, Mingers' Marshall correction splitting rule turns out to give lower error for most pruning rules and many different training set sizes, although by chance this is not the case reported in Table 3.

Finally, we reproduced Mingers' experimental method, using the pruning set for estimating error as well and obtained results roughly comparable to Mingers'. This indicates that the differences in the results for random splits arise largely from his use of the pruning set to determine errors.

## 6. Conclusion

The main differences between our results and those of Mingers can be summarized in two points:

- The random splitting rule performed very poorly on some of the data sets. When it did perform well, it was often because there was little difference between the base error (the error for a zero depth tree) and the optimum error. This differed from Minger's results because of our more careful experimental method.
- The Marshall correction is slightly better in error in some domains but significantly worse in others. The domains where it performs poorly are generally those where some classes have extreme proportions, not seen in Minger's data sets.

The result for the random splitting rule is to be expected. For instance, if only a few attributes for a domain are relevant to the class, splitting on other attributes partitions the data unnecessarily and eventually leads to the case where no reasonable classification can be made at the leaves. If however, all the attributes were more or less relevant, as for the digit LED domain, then random splitting does not perform too badly. Finally, if little gain over base error is achieved by the best splitting rule, then the random splitting rule will again be comparable in its error rate.

The results of using Mingers' Marshall correction indicated that the idea of favoring more equal partitions is a potentially promising approach, but it still has problems in some cases. Consider the four partitions given in Table 5. Under each counting table is the value of information gain (IG) and the value of information gain adjusted with the Marshall correction (IGM). In the left two tables, where the classes are fairly evenly distributed (71 versus 129), the two partitions have a similar information gain. The Marshall correction makes one prefer the bottom partition with the more even split. In the top left table, the sub-partition with size 190 surely has a similar class distribution to the original 200, so choosing this partition achieves very little for the majority of the data. In the right two tables where the classes are unevenly distributed (10 versus 190), the top right partition has a better information gain. This makes sense because it is an almost perfect partition. The Marshall correction makes one prefer the bottom partition, however. Because the classes are unevenly

*Table 5.* Counting tables for mild vs. extreme class counts.

| | Class | | |
|---|---|---|---|
| Test | Yes | No | Total |
| outcome-1 | 2 | 8 | 10 |
| outcome-2 | 69 | 121 | 190 |
| Total | 71 | 129 | 200 |

IG = 0.003, IGM = 0.0006

| | Class | | |
|---|---|---|---|
| Test | Yes | No | Total |
| outcome-1 | 2 | 8 | 10 |
| outcome-2 | 188 | 3 | 190 |
| Total | 190 | 10 | 200 |

IG = 0.118, IGM = 0.022

| | Class | | |
|---|---|---|---|
| Test | Yes | No | Total |
| outcome-1 | 18 | 24 | 42 |
| outcome-2 | 53 | 105 | 158 |
| Total | 71 | 129 | 200 |

IG = 0.003, IGM = 0.002

| | Class | | |
|---|---|---|---|
| Test | Yes | No | Total |
| outcome-1 | 42 | 8 | 50 |
| outcome-2 | 148 | 2 | 150 |
| Total | 190 | 10 | 200 |

IG = 0.035, IGM = 0.027

distributed, the sub-partition with size 190 will not have a similar class distribution to the original 200, so the Marshall correction is not appropriate in this case.

The disparity between our results and Mingers' demonstrates the care that must be taken when performing quantitative comparisons of different learning algorithms. We conclude that a random splitting rule performs significantly worse than the other measures. Finally, while not doubting that significant results can be obtained by comparing the performance of algorithms over two or three domains, we would suggest that, given the current availability of archived data sets, a more thorough and informative evaluation is provided by comparing performance over a larger number of varied domains.

## Notes

1. These data sets can be obtained from the authors on request.
2. The versions we used do not have real-valued attributes quantized into discrete attributes.

## References

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

Buntine, W. (1989). Learning classification rules using Bayes. In *Proceedings of the Sixth International Machine Learning Workshop*, Cornell, New York. Morgan Kaufmann.

Cestnik, B., Kononenko, I., & Bratko, I. (1987). Assistant86: A knowledge-elicitation tool for sophisticated users. In I. Bratko, & N. Lavrač (Eds.), *Progress in machine learning: Proceedings of EWSL-87*, (pp. 31–45), Bled, Yugoslavia. Sigma Press.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, *3*, 261–283.

Fisher, D., & McKusick, K. (1989). An empirical comparison of ID3 and back-propagation and machine learning classification methods. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, (pp. 788–793). Detroit: Morgan Kaufmann.

Fisher, R. (1936). Multiple measurements in taxonomic problems. *Annals of Eugenics*, VII, 179–188.

Kibler, D., & Langley, P. (1988). Machine learning as an experimental science. In D. Sleeman (Ed.), *Proceedings of the Third European Working Session on Learning*, (pp. 81–92). Glasgow: Pitman Publishing.

Mingers, J. (1989). An empirical comparison of selection measures for decision-tree induction. *Machine Learning*, *3*, 319–342.

Mooney, R., Shavlik, J., Towell, G., & Gove, A. (1989). An experimental comparison of symbolic and connectionist learning algorithms. In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, (pp. 775–780). Detroit: Morgan Kaufmann.

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

Quinlan, J. (1988). Simplifying decision trees. In B. Gaines & J. Boose (Eds.), *Knowledge acquisition for knowledge-based systems*, (pp. 239–252). London: Academic Press.

Quinlan, J. R. (1989). Unknown attribute values in induction. In *Proceedings of the Sixth International Machine Learning Workshop*. Cornell, New York: Morgan Kaufmann.

Quinlan, J. R., Compton, P., Horn, K., & Lazarus, L. (1987). Inductive knowledge acquisition: A case study. In J.R. Quinlan (Ed.), *Applications of expert systems*. London: Addison Wesley.

Schlimmer, J., & Granger Jr., R. (1986). Incremental learning from noisy data. *Machine Learning*, *1*, 317–354.

Weiss, S., & Kapouleas, I. (1989). An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, (pp. 781–787). Detroit: Morgan Kaufmann.