Bivariate Decision Trees

Jan C. Bioch, Onno van der Meer, and Rob Potharst

Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Abstract. Decision tree methods constitute an important and much used technique for classification problems. When such trees are used in a Datamining and Knowledge Discovery context, ease of interpretation of the resulting trees is an important requirement to be met. Decision trees with tests based on a single variable, as produced by methods such as ID3, C4.5 etc., often require a large number of tests to achieve an acceptable accuracy. This makes interpretation of these trees, which is an important reason for their use, disputable. Recently, a number of methods for constructing decision trees with multivariate tests have been presented. Multivariate decision trees are often smaller and more accurate than univariate trees; however, the use of linear combinations of the variables may result in trees that are hard to interpret. In this paper we consider trees with test bases on combinations of at most two variables. We show that bivariate decision trees are an interesting alternative to both uni– and multivariate trees, especially qua ease of interpretation.

1 Introduction

For classification problems decision tree methods such as ID3 and C4.5 [11] continue to be very popular. Solving a classification problem can be viewed as the search for a functional relationship, where the dependent variable represents the classes and the independent variables are the attributes of the objects to be classified [6,15]. Decision trees are also used in KDD as a method for the sub-process of model selection [2]. One drawback is the formidable size of many of these trees, even after pruning, in all but the most toy-like practical situations. To attain the desired accuracy, often dozens, sometimes hundreds of tests are needed, resulting in a decision tree, which is besides tedious to handle far from easily interpreted. To be sure, one of the main reasons why decision tree methods are preferred to competitors, such as neural networks, is supposedly their ease of interpretation. This ease of interpretation is paramount in a datamining context where variables and data items are abundant. One way to get smaller but equally accurate trees is to base the test in a node of the tree not on a single variable, but on a number of variables [4]. Up until now, most implementations of this idea, usually called multivariate decision trees, make use of tests, which gauge the value of a linear combination of many variables per node. Though the resulting trees become much smaller, they are equally hard to interpret, as anyone knows, who has ever faced the problem of interpreting the conjunction of a number of tests like

$$1.3450x_1 - 0.6333x_2 - 0.2214x_3 + 0.8779x_4 + 0.5667x_5 - 1.1432x_6 \leq 0.8883 \quad (1)$$

or worse. In this paper we investigate the possibility of restricting the number of variables per test to two, thus preventing hardships like (1) while at the same time we hope to obtain sufficiently accurate and not too large trees. Bivariate methods have also been used in the KDD area, among others by [15]. In section 2 we start with describing both univariate and multivariate decision tree methods. In section 3 two variants of our own proposed method are described. The next two sections apply our methods, first to a constructed dataset, and subsequently to five well known real life data sets. In the last section we conclude that our so-called bivariate decision trees can be competitive with multivariate trees in both size and accuracy, without sacrificing the interpretability of the resulting trees. Our second method, which we call BIT2, produces very elegant, small decision trees and is computationally cheap. More details can be found in [1].

2 Inducing Binary Decision Trees

In this paper we will study tree classifiers (see e.g. [12]) for multiclass classification problems. The trees we will consider consist of nodes, leaves and branches, e.g.

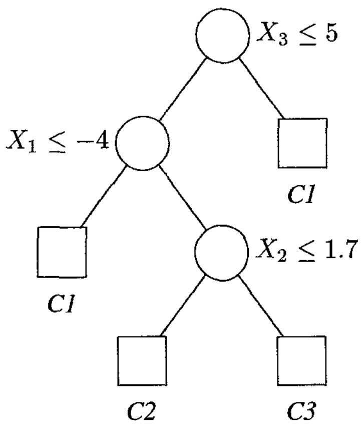

Fig. 1. Binary Decision Tree: Example

With each node a test is associated, which in general will be of the form

$$\sum_{i=1}^{n} a_i X_i \leq T. \tag{2}$$

Here $X_1, \ldots, X_n$ are features, the a_i are real coefficients and T is a real cut-off point. With each leaf one of the classes $C_1, \ldots, C_k$ is associated. Furthermore, all splits will be *binary*: go left if the test is satisfied, and right if not. Thus, the example tree of Figure 1 will classify an example as C_1, if $X_3 \leq 5$ and $X_1 \leq -4$ or if $X_3 > 5$. For the induction of such a decision tree, we will make use of a training set of examples or data points, for each of which both the values of all the features, and a class label are given.

2.1 Univariate Trees

Univariate decision trees recursively partition the training set by finding tests of the form $X_i \leq T$ that minimize an impurity measure. Using this test the data is partitioned into two groups (left/right) according to whether the data points satisfy the inequality.

As an example consider the following (sorted) six values for X and corresponding class labels:

	1	2	3	4	5	6
X	-0.9	-0.5	-0.1	0.4	0.7	0.9
Class	1	1	2	1	2	2

The possible values for the threshold T are the midpoints between two adjacent values of X. If we use a convex splitting criterion, we know from the work of Fayyad and Irani [5] that we need only consider midpoints between values of X with different class labels. For instance, (0.4+0.7)/2 is a possible threshold and leads to the following partition:

	Class 1	Class 2	Total
Left	3	1	4
Right	0	2	2

There are many splitting criteria available, however, we limit our attention to the Gini Index as proposed by Breiman *et al.* [3] in the form as given by Murthy [8] for k classes:

$$I = \left(N_l(1 - \sum_{i=1}^{k}(L_i/N_l)^2) + N_r(1 - \sum_{i=1}^{k}(R_i/N_r)^2) \right) /N,$$

where $L_i(R_i)$ is the number of examples of class i in the left (right) node, and $N_l(N_r)$ is the total number of examples in the left (right) node. Calculating the Gini Index for the example gives

$$\left(4(1.0 - (3/4)^2 - (1/4)^2) + 2(1.0 - (0/2)^2 - (2/2)^2)\right)/6.$$

2.2 Multivariate Trees

Several methods for constructing multivariate decision trees exist. These include CART [3], Recursive Least Squares [4], Linear Machine Decision Trees [14] and OC1 [8,9]. Of these only the first and last produce binary trees for any number of classes, so we confine our discussion to CART and its descendant OC1.

2.3 CART

In their epoch making 1984 book *Classification And Regression Trees*, abbreviated CART, Breiman *et al.* [3] describe both univariate and multivariate classification trees. The algorithm they propose for the multivariate case searches in each node for the linear test, which splits the data for that node into two subsets showing the least impurity. For a given node, this is done as follows:

1. Normalize the data by centering all the data points for the node at their medians, and dividing by their interquartile ranges; we then get N data points $x^{(1)}, \ldots, x^{(N)}$ with $x^{(j)} = (x_1^{(j)}, \ldots, x_n^{(j)})$ a vector of values for each of the n features.

2. Find the best univariate split $X_i \leq T$, in the way as descibed above.

3. Next, in a number of cycles, we try to improve upon a given split $V = \sum_{i=1}^{n} a_i X_i \leq T$ by successively considering splits of the form:

$$V - \delta(X_i + \gamma) \leq T. \tag{3}$$

A complete cycle consists of all updates of form (3). For instance, for $i = 1$, we can find the best δ by calculating

$$u^{(j)} = \frac{v^{(j)} - T}{x_1^{(j)} + \gamma}$$

for $j = 1, \ldots, N$, taking γ as a fixed value. By considering the midpoints of the ordered $u^{(j)}$ as possible candidates for δ, we can calculate the best split 3. It is proposed, that this procedure is performed for three different values of γ, namely $-0.25, 0$ and 0.25. The best of these three is then taken as the δ and γ that are used to update V to V' as follows:

$$V' = \sum_{i=1}^{n} a'_i X_i$$

with

$$a'_i = \begin{cases} a_i & \text{for } i \neq 1 \\ a_1 - \delta & \text{otherwise} \end{cases}$$

and

$$T' = T + \delta\gamma.$$

This step is repeated subsequently for the other variables $X_2, \ldots, X_n$. If, at the end of such a cycle we have obtained test $V = \sum a_i X_i \leq T$, we next try to improve T by finding the best test $\sum a_i X_i \leq T$, fixing the a_i and changing only the threshold T.

4. We stop cycling when the decrease in impurity from one cycle to the next is below a predetermined small level ϵ.

This algorithm aims at finding a global minimum for the impurity of the considered linear splits. However, the algorithm can get stuck in a local minimum.

2.4 OC1

The linear combination search algorithm of CART, as described in the previous subsection, may not always get close to a globally optimal split. This problem has been addressed by Murthy [8] in the OC1 method. The parameter update method of OC1 follows the CART method, but includes random perturbations of the parameters when a local minimum is reached and restarts from random locations (the first starting location is the best univariate split).

3 Bivariate Decision Trees

With multivariate decision trees, each node contains a test of the form (2). In order to enhance interpretability, we want to reduce the number of variables used in (2) to at most two, making it a test of the form

$$a_i X_i + a_j X_j \leq T. \tag{4}$$

Of course, in different nodes, different choices for variables X_i and X_j may be used. Our proposed method has two variants, which we will call BIT1 and BIT2. In the first variant, in each node we look for the best test of type (4) allowing all combinations of X_i and X_j, and all possible coefficients a_i and a_j. The second variant only allows certain values of a_i and a_j: we look for the best test of the type

$$X_i + X_j \leq T \text{ or } X_i - X_j \leq T.$$

Although it may look like this restricts the set of admissable tests, the input space partition induced by a number of such tests is still much more flexible than in the univariate case, as can be seen from Figure 2. Both variants only use a bivarate test if it improves upon the best univariate test.

3.1 BIT1

For each combination of attributes X_i and X_j of the form $a_i X_i + a_j X_j \leq T$ we use the CART method of parameter updating to find values for a_i, a_j and T. This process is started by setting the parameters to values corresponding to the univariate test for attribute X_i. For instance, if the best univariate test for X_2 is $X_2 \leq 4$, and we combine X_2 and X_3, we set $a_2 = 1$, $a_3 = 0$ and $T = 4$.

If we have n variables we need to look at $n(n-1)$ combinations:

```
1 Find best univariate split
2 Find best bivariate split:
  for $i = 1$ to $n$
      for $j = 1$ to $n$ and $j \neq i$
          find $a_i X_i + a_j X_j \leq T$
```

If the bivariate split has lower impurity than the univariate split, we use the bivariate split. Another possibility is to choose the bivariate split only if the improvement in impurity exceeds a threshold, say 10 percent; however, experiments suggest that this procedure often leads to increased tree size.

3.2 BIT2

A much faster method is to consider tests of the form $X_i + X_j \leq T$ or $X_i - X_j \leq T$, in addition to the best univariate test. Where univariate tests form lines parallel to the axis, these bivariate tests form 45 degree lines. Selecting the optimal line is analogous to finding the best univariate split. The combinations of X_i and X_j are easily calculated from the training data and give new univariate problems, so we can use the procedure based on the results of Fayyad and Irani described above. The procedure of finding the best test is similar to that of BIT1. Note that this method is not only faster than BIT1, it also leads to simpler tests. Furthermore, it is flexible in modeling class configurations in input space.

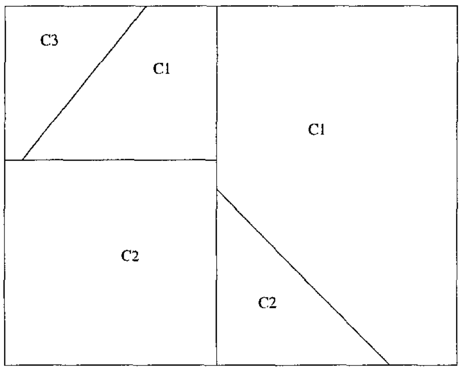

Fig. 2. Input space partition by a BIT2 tree

This method assumes that all attributes are on the same scale. If one attribute takes on values in $[1, \ldots, 2]$ and a second values from $[3, \ldots, 187]$, the combination of the attributes will have no effect. This can be solved by scaling all attributes to have the same range, e.g. $[0, 1]$ and adjusting the solution for the original values. For instance, if we find the test $X_1 + X_2 \leq 1.5$ for the scaled data, the test on the original data is

$$\frac{X_1 - 1}{2 - 1} + \frac{X_2 - 3}{187 - 3} \leq 1.5$$

or $X_1 + 0.00543X_2 \leq 2.5163$.

4 Capturing the model

In this section we will investigate whether our proposed bivariate decision tree method is able to recover an underlying model from an artificial dataset, in which interactions between two variables occur. We will test our method on two artificial data sets. The first is the monk1 data set [13]. This data set contains 6 attributes and two classes. The rule that generates the data is: if ($x_1 = x_2$ or $x_5 = 1$) then *yes* else *no*. Note that x_5 takes on values form $[1, 2, 3, 4]$ and x_1 and x_2 from $[1, 2, 3]$. The bivariate methods produce the trees shown in figure 3, which capture the rule perfectly. In the BIT2 tree, the first node represents the $x_5 = 1$ condition as $x_5 \leq 2$, and the $x_1 = x_2$ condition requires two tests.

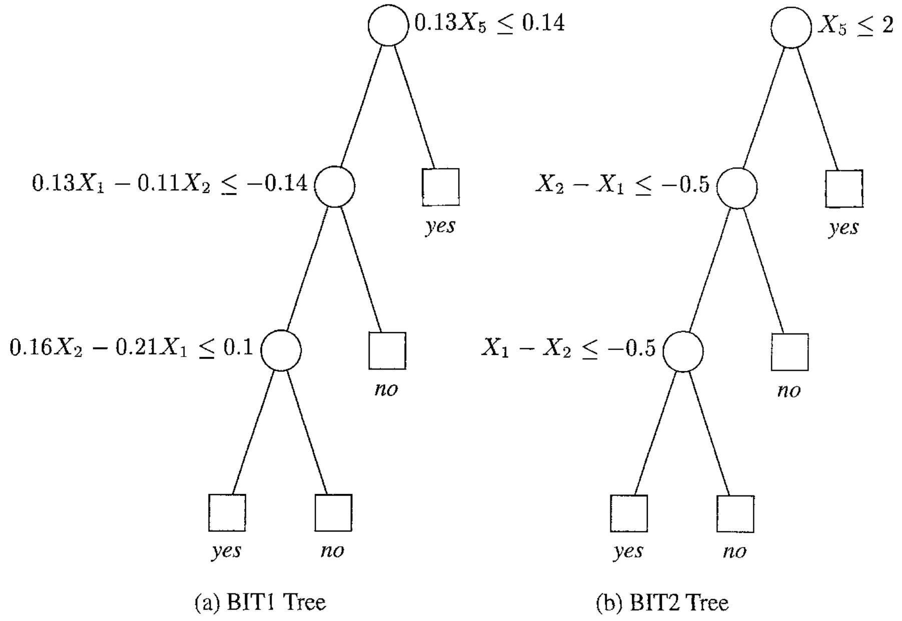

Fig. 3. Trees for the monk1 example

As a second example we constructed a data set again with six attributes and two classes. The rule that generates the data is: if $(x_1 \leq x_2)$ and $(x_4 \leq x_6)$ then *yes* else *no*. A univariate decision tree approximates this rule with a series of orthogonal splits, resulting in a very large tree containing about 40 leaves. The tree produced by OC1, on the other hand, contains only one test:

$$1.3450x_1 - 0.6333x_2 - 0.2214x_3 + 0.8779x_4 + 0.5667x_5 - 1.1432x_6 \leq 0.8883.$$

It is hard to see the structure of the problem in this representation. The bivariate methods can represent this problem exactly with only two tests, and do not require pruning. Notice that the tree resulting from the BIT2 method is more accurate than both the BIT1 tree and the multivariate tree produced by OC1. Figure 4 shows the trees produced by the bivariate methods.

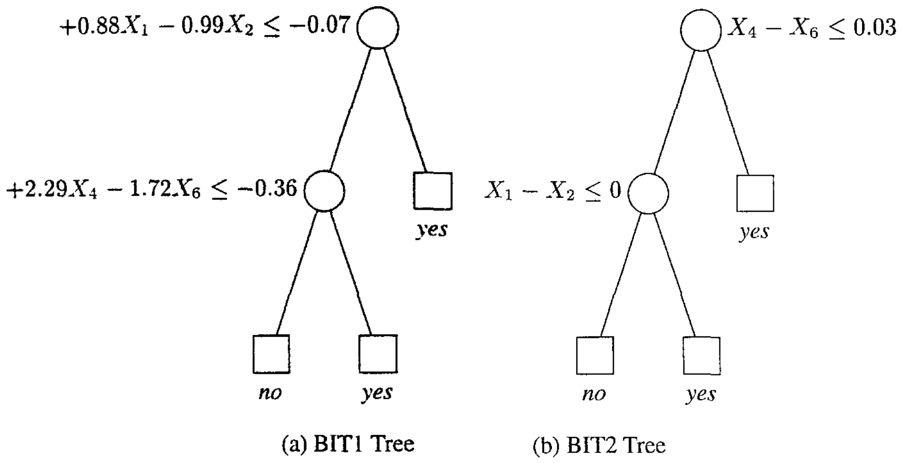

Fig. 4. Trees for the second example

5 Experiments

5.1 Experimental Setup

We used the following experimental setup for the bivariate tree methods BIT1 and BIT2, and OC1 (uni– and multivariate). We performed ten complete tenfold cross validations for each data set. In each tenfold cross validation the data was split into ten disjoint partitions of 90 percent training data and ten percent testing data, with each method using the same data. The final results are the mean values obtained by averaging over ten such cross validations.

We set up the methods mentioned to use the same pruning procedure and the same cross validation partitions of the data. The pruning procedure was cost–complexity pruning with the 1-Standard Error rule [3] using ten percent of the training data. Again, the data used for pruning was identical for all four methods. All methods used the Gini Index as the impurity measure. All attributes in the datasets were scaled to a range of $[0, 1]$.

A further comparison was made with the results from a similar cross–validation experiment as given by Quinlan [11] for C4.5 release 8.

5.2 Description of the Datasets

This section briefly describes the five data sets used in the experiments (see table 1). All of them are well-known; for instance they are among the datasets used in [11]. The datasets were obtained from the UCI Repository of Machine Learning Databases [7].

1. Glass data: the aim here is to correctly identify a piece of glass as belonging to one of six possible classes, based on the refractive index and chemical composition of the glass.

2. Diabetes data: contains records of signs of diabetes in Pima Indian females over 21 years based on eight continuous attributes.
3. Breast cancer: the data was compiled by Dr. William H. Wolberg at the University of Wisconsin Hospitals [10] and has nine attributes and a diagnosis of the tumor as benign (65.5% of the cases) or malignant (34.5%).
4. Heart disease data: tests presence (44.4% of the cases) or absence of heart disease based on 13 patient characteristics.
5. Wave data: this data set was constructed by Breiman et al [3] as an example, and has 21 variables and three classes and is quite difficult to learn.

name	cases	attr	classes
glass	214	9	6
diabetes(pima)	768	8	2
breast cancer	699	9	2
heart	270	13	2
wave	300	21	3

Table 1. Summary of the Datasets

5.3 Results

The results from the experiments are summarized in table 2 for the pruned trees and table 3 for the unpruned trees. The tables give the mean accuracy and mean tree size from ten cross–validation runs, together with the standard error. From these table we can conclude that the bivariate methods often perform much better than univariate trees, while the size is usually quite close to the multivariate trees. The BIT2 methods only performs slightly worse than the BIT1 method on the glass data. As the BIT1 method is computationally much more expensive, the BIT2 method would be more useful in practice.

From additional experiments we can note the effect of scaling on the BIT2 method. Of the five datasets, only the cancer and wave data have all attributes on the same scale, making scaling redundant. On the diabetes data difference between scaled and original data was not significant, but on the heart data the scaling resulted in much smaller trees (4 leaves vs 11 leaves) and better performance than the original data. On the glass data, however, the scaling resulted in decreased performance. This could explain the rather low accuracy for this data.

6 Conclusion

We have discussed two methods of constructing bivariate decision trees, in which per test node at most two attributes are used. We have shown that bivariate tree methods can improve the interpretability of the resulting trees (both small trees and simple tests),

method	glass	diabetes	cancer	heart	wave
BIT1	65.3±1.1	74.3±0.7	95.4±0.3	78.5±0.3	76.1±1.3
	6.2±2.1	5.2±2.5	2.8±0.2	4.1±0.5	5.0±1.6
BIT2	64.8±0.9	74.7±0.7	95.4±0.3	78.5±0.3	76.2±1.0
	6.8±2.5	5.8±2.7	2.6±0.2	4.1±0.5	5.2±1.3
ocl - uni	63.8±2.8	73.8±1.1	94.5±0.8	75.9±0.8	69.1±1.8
	8.1±4.5	8.4±6.3	6.4±2.2	5.2±1.1	4.8±2.1
ocl - mlt	64.0±0.9	74.2±0.8	95.0±0.5	77.4±0.4	78.2±1.4
	7.9±2.1	4.3±3.6	2.8±0.9	3.9±0.7	3.8±0.9
c4.5	68.2±0.8	74.6±0.3	94.7±0.2	77.0±0.2	72.7±0.3
	45.7±0.4	44.0±1.6	25.0±0.5	19.1±0.6	44.6±0.4

Table 2. Accuracy and Size for Pruned Trees

	glass	diabetes	cancer	heart	wave
BIT1	65.9±1.3	70.9±0.9	94.4±0.5	73.5±0.5	68.6±1.8
	24.3±2.1	75.3±1.2	8.2±0.3	15.1±0.6	14.9±0.8
BIT2	63.3±1.3	70.2±1.4	93.1±0.4	73.0±0.4	69.0±2.1
	27.8±2.4	75.5±2.3	11.1±0.3	19.9±0.5	15.1±1.1
ocl - uni	66.8±3.0	70.5±4.3	93.9±0.9	71.5±0.8	68.6±2.7
	44.6±4.8	122.4±8.5	20.5±2.1	41.9±0.9	36.3±1.3
ocl - mlt	66.3±1.5	69.9±1.2	91.2±1.1	79.3±0.5	71.0±1.2
	28.4±1.3	50.3±4.2	16.1±0.8	21.2±0.8	24.2±0.8
c4.5	67.5±0.9	74.2±0.4	94.2±0.2	77.6±0.2	72.6±0.3
	45.5±0.4	52.3±1.3	41.4±0.3	48.4±0.5	45.4±0.4

Table 3. Accuracy and Size for Unpruned Trees

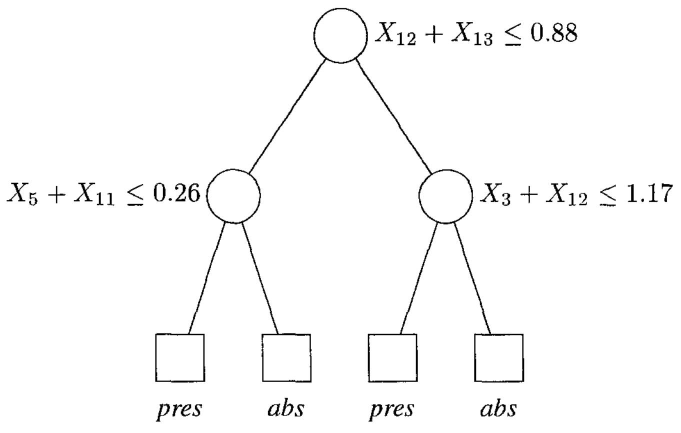

Fig. 5. BIT2 tree for the scaled heart data

with accuracy at least as good as C4.5. We found both the BIT1 and the BIT2 method to give good results. In addition, despite its simplicity BIT2 proved to be quite effective. Topics for further research: we are currently investigating the issue of scaling for the BIT2 method. This method could also be extended to include interaction terms of the form $X_i X_j$.

References

1. J.C. Bioch, O.R. van der Meer, R. Potharst. Bivariate Decision Trees. Technical Report, Department of Computer Science, Erasmus University Rotterdam, 1996.
2. R.J. Brachman. The Process of Knowledge Discovery in Databases, in: U.M. Fayyad *et al.* eds. *Advances in Knowledge Discovery and Datamining*, Ch.2, AAAI/MIT Press, Cambridge 1996.
3. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth (reprinted by Chapman & Hall), 1984.
4. C.E. Brodley and P.E. Utgoff. Multivariate decision trees. *Machine Learning*, 19: 45–77, 1995.
5. U.M. Fayyad and K.B. Irani. Multi–interval discretization of continous–valued attributes for classification learning. In *Proceedings Thirteenth International Joint Conference on Artificial Intelligence*, pages 1022–1027. San Fransisco: Morgan Kaufmann, 1993.
6. B. Koehn and J.M. Zytkow. Experimenting and Theorizing in Theory Formation, in: Z. Ras and M. Zemankova eds. *Proceedings of the International Symposium of Methodologies for Intelligent Systems*, ACM SIGART Press, 296–307, 1986.
7. P.M. Murphy and D.W. Aha. UCI repository of machine learning databases and domain theories. Technical report, University of California, 1992.
Web Page: `http://www.ics.uci.edu/~mlearn/MLRepository.html`.
8. S. Murthy, S. Kasif, S. Salzberg, and R. Beigel. OC1: Randomized induction of oblique decision trees. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 322–327, 1993.
9. S. Murthy. *On Growing Better Decision Trees from Data.* PhD thesis, John Hopkins University, Baltimore, 1995.
10. O.L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming. *SIAM News*, 23(5):1–18, 1990.
11. J.R. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.
12. B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.
13. S. Thrun. Comparison of machine learning algorithms. Technical Report CMU-CS-91-197, Carnegie Mellon University, Department of Computer Science, 1991.
14. P.E. Utgoff and C.E. Brodley. Linear machine decision trees. Technical Report COINS Technical Report 91–10, University of Massachusetts, Department of Computer and Information Science, 1991.
15. J.M. Zytkow. Automated Discovery of Empirical Laws. *Fundamenta Informaticae*, 27: 299–318, 1996.