Machine Learning, 6, 81-92 (1991)

Technical Note
A Distance-Based Attribute Selection Measure for Decision Tree Induction

R. LÓPEZ DE MÁNTARAS (MANTARAS@CEAB.ES)
Centre of Advanced Studies, CSIC, 17300 Blanes, Girona, Spain

Editor: J. Carbonell

Abstract. This note introduces a new attribute selection measure for ID3-like inductive algorithms. This measure is based on a distance between partitions such that the selected attribute in a node induces the partition which is closest to the correct partition of the subset of training examples corresponding to this node. The relationship of this measure with Quinlan's information gain is also established. It is also formally proved that our distance is not biased towards attributes with large numbers of values. Experimental studies with this distance confirm previously reported results showing that the predictive accuracy of induced decision trees is not sensitive to the goodness of the attribute selection measure. However, this distance produces smaller trees than the gain ratio measure of Quinlan, especially in the case of data whose attributes have significantly different numbers of values.

Keywords. Distance between partitions, decision tree induction, information measures

1. Introduction

ID3 (Quinlan, 1979, 1986) is a well-known inductive learning algorithm to induce classification rules in the form of a decision tree. ID3 works on a set of examples described in terms of "attribute-value" pairs. Each attribute measures some feature of an object by means of a value among a set of discrete, mutually exclusive values. ID3 performs a heuristic hill-climbing search without backtracking through the space of possible decision trees. For each non-terminal node of the tree, ID3 recursively selects an attribute and creates a branch for each value of the attribute. Therefore, a fundamental step in this algorithm is the selection of the attribute at each node. Quinlan introduced a selection measure based on the computation of an information gain for each attribute and the attribute that maximizes this gain is selected. The selected attribute is the one that generates a partition in which the examples are distributed less randomly over the classes. This note starts by recalling in some detail Quinlan's information Gain. A notable disadvantage of this measure is that it is biased towards selecting attributes with many values. This motivated Quinlan to define the Gain Ratio which mitigates this bias but suffers from other disadvantages that we will describe. We introduce a *distance* between partitions as attribute selection measure and we formally prove that it is not biased towards many-valued attributes. The relation of the proposed distance with Quinlan's Gain is established and the advantages of our distance over Quinlan's Gain Ratio are shown.

2. Quinlan's information gain measure

Let X be a finite set of examples and $\{A_1, \ldots, A_p\}$ a set of attributes. For each attribute A_k, ID3 measures the information gained by branching on the values of attribute A_k using the following information Gain measure

$$Gain(A_k, X) = I(X) - E(A_k, X) \tag{1}$$

where

$$I(X) = -\sum_{j=1}^{m} P_j \, log_2 P_j , \qquad P_j = \frac{|X \cap F_j|}{|X|} \tag{2}$$

measures the randomness of the distribution of examples in X over m possible classes. P_j is the probability of occurrence of each class F_j in the set X of examples, defined as the proportion of examples in X that belong to class F_j, and $E(A_K, X)$ is given by

$$E(A_K, X) = \sum_{i=1}^{n} \frac{|X_i|}{|X|} I(X_i) \tag{3}$$

where:

- n is the number of possible values of attribute A_K.
- $|X_i|$ is the number of examples in X having value V_i for attribute A_K, and
- $|X|$ is the number of examples in the node

Note that the sets $X_1, \ldots, X_n$ form a partition on X generated by the n values of A_k. $I(X_i)$ measures the randomness of the distribution of examples in the set X_i, over the possible classes and is given by

$$I(X_i) = -\sum_{j=1}^{m} \frac{|X_i \cap F_j|}{|X_i|} \, log_2 \frac{|X_i \cap F_j|}{|X_i|} \tag{4}$$

$E(A_K, X)$ is, therefore, the expected information for the tree with A_K as root. This expected information is the weighted average, over the n values of attribute A_K, of the measures $I(X_i)$.

The attribute selected is the one that maximizes the above Gain. However, as has already been pointed out in the literature (Hart, 1984; Kononenko et al., 1984; Quinlan, 1986), this measure is biased in favor of attributes with a large number of values. Quinlan (1986) introduced a modification of the Gain measures to compensate for this bias. The modification consists in dividing $Gain(A_k, X)$ by the following expression

$$IV(A_k) = -\sum_{i=1}^{n} \frac{|X_i|}{|X|} \, log_2 \frac{|X_i|}{|X|} \tag{5}$$

obtaining the Gain Ratio

$$G_R(A_k, X) = \frac{I(X) - E(A_k, X)}{IV(A_k)} \tag{6}$$

$IV(A_k)$ measures the information content of the attribute itself, and according to Quinlan, "the rationale behind this is that as much as possible of the information provided by determining the value of an attribute should be useful for classification purpose." However, the modified Gain has the following limitations: it may not be always defined (the denominator may be zero), and it may choose attributes with very low $IV(A_k)$ rather than those with high gain. To avoid this Quinlan proposes to apply the Gain Ratio to select from among those attributes whose initial (not modified) Gain is at least as high as the average Gain of all the attributes.

Bratko and Kononenko (1986) and Breiman et al. (1984) take a different approach to this multivalued attributes problem by grouping the various attribute values together so that all the attributes become bi-valued. However, binarized trees have the problem of being more difficult to interpret.

In this paper we introduce a new attribute selection measure that provides a clearer and more formal framework for attribute selection and solves the problem of bias in favor of multivalued attributes without having the limitations of Quinlan's Gain Ratio.

3. An alternate selection criterion

Instead of using Quinlan's Gain, we propose an attribute selection criterion based on a distance between partitions. The chosen attribute in a node will be that whose corresponding partition is the closest (in terms of the distance) to the correct partition of the subset of examples in this node.

3.1. Distances between partitions

First let us recall some fundamental results of information theory.

Let us consider two partitions on the same set X; a partition P_A whose classes will be denoted A_i for $1 \leq i \leq$ n and a partition P_B, whose classes will be denoted B_j for $1 \leq j \leq$ m.

Let us consider the following probabilities

$$\begin{aligned} P_i &= P(A_i) \\ P_j &= P(B_j) \\ P_{ij} &= P(A_i \cap B_j) \\ P_{j/i} &= P(B_j/A_i) \end{aligned}$$

for all $1 \leq i \leq$ n and $1 \leq j \leq$ m.

The average information of partition P_A which measures the randomness of the distribution of elements of X over the n classes of the partition is

$$I(P_A) = - \sum_{i=1}^{n} P_i log_2 P_i \tag{7}$$

Similarly, for P_B is

$$I(P_B) = - \sum_{j=1}^{m} P_j log_2 P_j \tag{8}$$

Furthermore, the mutual average information of the intersection of two partitions $P_A \cap P_B$ is

$$I(P_A \cap P_B) = - \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} log_2 P_{ij} \tag{9}$$

and the conditional information of P_B given P_A is

$$I(P_B/P_A) = I(P_B \cap P_A) - I(P_A) = - \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} log_2 \left(\frac{P_{ij}}{P_i} \right) = - \sum_{i=1}^{n} P_i \sum_{j=1}^{m} P_{j/i} log_2 P_{j/i} \tag{10}$$

Now we can introduce two distances between partitions, one being a normalization of the other. We will show a relationship between the normalized distance and Quinlan's Gain.

Proposition-1

The measure $d(P_A, P_B) = I(P_B/P_A) + I(P_A/P_B)$ is a metric distance measure (López de Mántaras, 1977), that is, for any partitions P_A, P_B, and P_C on X it satisfies

(i) $d(P_A, P_B) \geq 0$ and the equality holds iff $P_A = P_B$ (11a)
(ii) $d(P_A, P_B) = d(P_B, P_A)$ (11b)
(iii) $d(P_A, P_B) + d(P_A, P_C) \geq d(P_B, P_C)$ (11c)

Proof:

Properties (i) and (ii) are trivial. Let us prove the triangular inequality (iii). Let us first show the following inequality

$$I(P_B/P_A) + I(P_A/P_C) \geq I(P_B/P_C) \tag{12}$$

Since $I(P_B/P_A) \leq I(P_B)$, we can write

$$I(P_B/P_A) + I(P_A/P_C) \geq I(P_B/(P_A \cap P_C)) + I(P_A/P_C) \quad (13)$$

Now by (10) we have

$$I(P_B/(P_A \cap P_C)) + I(P_A/P_C) = I((P_B \cap P_A)/P_C) \quad (14)$$

On the other hand, by (10) we know that

$$I((P_B \cap P_A)/P_C) \geq I(P_B/P_C) \quad (15)$$

combining (13), (14) and (15) we have

$$I(P_B/P_A) + I(P_A/P_C) \geq I(P_B/P_C) \quad (16)$$

Similarly permuting P_B and P_C in (16) we obtain

$$I(P_C/P_A) + I(P_A/P_B) \geq I(P_C/P_B) \quad (17)$$

Finally, adding (16) and (17) we obtain

$$I(P_B/P_A) + I(P_A/P_B) + I(P_A/P_C) + I(P_C/P_A) \geq I(P_B/P_C) + I(P_C/P_B)$$

that is: $d(P_B, P_A) + d(P_A, P_C) \geq d(P_B, P_C)$.

Proposition 2

The normalization

$$d_N(P_A, P_B) = \frac{d(P_A, P_B)}{I(P_A \cap P_B)}$$

is a distance in [0, 1] (López de Mántaras, 1977).

Proof:

Properties (i) and (ii) are clearly preserved. In order to prove that the triangular inequality (iii) also holds, let us first prove the following inequality

$$\frac{I(P_B/P_A)}{I(P_B \cap P_A)} + \frac{I(P_A/P_C)}{I(P_A \cap P_C)} \geq \frac{I(P_B/P_C)}{I(P_B \cap P_C)} \quad (18)$$

from (10) we have:

$$\frac{I(P_B/P_A)}{I(P_B \cap P_A)} + \frac{I(P_A/P_C)}{I(P_A \cap P_C)} = \frac{I(P_B/P_A)}{I(P_B/P_A) + I(P_A)} + \frac{I(P_A/P_C)}{I(P_A/P_C) + I(P_C)} \geq$$

$$\geq \frac{I(P_B/P_A)}{I(P_B/P_A) + I(P_A/P_C) + I(P_C)} + \frac{I(P_A/P_C)}{I(P_B/P_A) + I(P_A/P_C) + I(P_C)} = \frac{I(P_B/P_A) + I(P_A/P_C)}{I(P_B/P_A) + I(P_A/P_C) + I(P_C)}$$

$$\geq \frac{I(P_B/P_C)}{I(P_B/P_C) + I(P_C)} = \frac{I(P_B/P_C)}{I(P_B \cap P_C)}$$

That is (18) is true.

Now permuting P_B and P_C in (18) we have

$$\frac{I(P_C/P_A)}{I(P_C \cap P_A)} + \frac{I(P_A/P_B)}{I(P_A \cap P_B)} \geq \frac{I(P_C/P_B)}{I(P_C \cap P_B)} \tag{19}$$

Finally, adding (18) and (19) we obtain

$$\frac{I(P_B/P_A) + I(P_A/P_B)}{I(P_A \cap P_B)} + \frac{I(P_A/P_C) + I(P_C/P_A)}{I(P_A \cap P_C)} \geq \frac{I(P_B/P_C) + I(P_C/P_B)}{I(P_B \cap P_C)} \tag{20}$$

Therefore, the triangular inequality also holds.

Finally, let us prove that $d_N(P_B, P_A) \in [0, 1]$

We have that $I(P_B/P_A) = I(P_B \cap P_A) - I(P_A)$

and $I(P_A/P_B) = I(P_B \cap P_A) - I(P_B)$

Then

$$d_N(P_A, P_B) = \frac{I(P_B/P_A) + I(P_A/P_B)}{I(P_B \cap P_A)} = 2 - \frac{I(P_A) + I(P_B)}{I(P_B \cap P_A)}$$

but

$$1 \leq \frac{I(P_A) + I(P_B)}{I(P_B \cap P_A)} \leq 2$$

because from $I(P_B/P_A) \leq I(P_B)$ we have

$$I(P_B \cap P_A) \leq I(P_B) + I(P_A)$$

and because we also have

$$I(P_B \cap P_A) \geq I(P_A)$$

and $I(P_B \cap P_A) \geq I(P_B)$.

Therefore $2 \times I(P_B \cap P_A) \geq I(P_B) + I(P_A)$

4. Relation with Quinlan's information gain

Let us first reformulate Quinlan's Gain in terms of measures of information on partitions in order to see the relationship with the proposed distance. Let P_C be the partition $\{C_1, \ldots, C_m\}$ of the set X of examples in its m classes, and let P_V be the partition $\{X_1, \ldots, X_n\}$ generated by the n possible values of attribute A_k (see Paragraph 2)

It is easy to check that the expression $I(X)$ in Quinlan's Gain is the average information of partition P_C as defined in section 3.1 above. That is

$$I(X) = I(P_C) = -\sum_{j=1}^{m} P_j log_2 P_j \tag{21}$$

On the other hand, Expression (3) can be rewritten as follows

$$E(A_K, X) = -\sum_{i=1}^{n} P_i \sum_{j=1}^{m} P_{j/i} log_2 P_{j/i} \tag{22}$$

where

$$P_{j/i} = \frac{|X_i \cap C_j|}{|X_i|}$$

and

$$P_i = \frac{|X_i|}{|X|}$$

but (22) is the conditional information of P_C given P_V. Therefore, Quinlan's Gain can be expressed, in terms of measures of information on partitions, as follows

$$Gain(A_K, X) = I(P_C) - I(P_C/P_V) \tag{23}$$

Once we have expressed Quinlan's Gain in such terms, it is easy to see its relationship with our normalized distance:

Adding and subtracting $I(P_V/P_C)$ to (23) we have

$$\begin{aligned} Gain(A_K, X) &= I(P_V/P_C) + I(P_C) - I(P_C/P_V) - I(P_V/P_C) = \\ &= I(P_V \cap P_C) - I(P_C/P_V) - I(P_V/P_C) \end{aligned}$$

Now, dividing by $I(P_V \cap P_C)$ we obtain

$$\frac{Gain(A_K, X)}{I(P_V \cap P_C)} = 1 - \frac{I(P_C/P_V) + I(P_V/P_C)}{I(P_V \cap P_C)}$$

We have then

$$1 - d_N(P_C, P_V) = \frac{Gain(A_K, X)}{I(P_V \cap P_C)} \in [0, 1] \tag{24}$$

That is, mathematically speaking, Quinlan's Gain normalized by the mutual information of P_C and P_V, is a similarity relation.

Furthermore, Quinlan's Gain Ratio can also be expressed in terms of information measures on partitions as follows

$$G_R(A_K, X) = \frac{I(P_C) - (P_C/P_V)}{I(P_V)} = \frac{Gain(A_K, X)}{I(P_V)} \tag{25}$$

since

$$IV(A_K) = -\sum_{i=1}^{n} \frac{|X_i|}{|X|} log_2 \frac{|X_i|}{|X|} = -\sum_{i=1}^{n} P_i log_2 P_i = I(P_V)$$

We notice that the difference between (24) and (25) is that "$1 - d_N(P_C, P_V)$" is equivalent to normalizing Quinlan's gain by the mutual information $I(P_V \cap P_C)$ instead of the information $I(P_V)$ associated with the partition generated by attribute A_K. It is interesting to notice that in our case, $I(P_C \cap P_V)$ cannot be zero if the numerator is different from zero contrarily to the Gain ratio expression which may not always be defined. Furthermore, our normalization also solves the problem of choosing attributes with very low information content $I(P_V)$ rather than with high Gain because we always have $I(P_V \cap P_C) > Gain(A_K, X)$. Therefore, instead of selecting the attribute that maximizes Quinlan's Gain ratio, we propose to select the attribute that minimizes our normalized distance.

Quinlan empirically found that his Gain Ratio criterion is efficient in compensating the bias in favor of attributes with larger number of values. With the proposed distance this is also true and, furthermore, it can be formally proved. In order to prove it let us first recall Quinlan's analysis concerning the bias of his gain (Quinlan, 1986): Let A be an attribute with values $A_1, A_2, \ldots, A_n$ and let A' be an attribute constructed from A by splitting one of its n values into two. (The partition P'_V generated by A' is finer than the partition P_V generated by A, that is $P'_V \subset P_V$). In this case, *if the values of A were sufficiently fine* for the induction task at hand, we would not expect this refinement to increase the usefulness of A'. Rather, as Quinlan writes, it might be anticipated that excessive fineness would tend to obscure structure in the training set so that A' should be in fact less useful than A. However it can be proved that $Gain(A', X)$ is greater than $Gain(A, X)$ with the result that A' would be selected. With the proposed distance this is not the case as the following theorem shows.

Theorem

Let P_C, P_V and P'_V be partitions on the same set X such that P'_V is finer than P_V and let us assume that all the examples in X_k of P_V belong to C_l of P_C. Then we have that

$$d(P_V, P_C) \leq d(P'_V, P_C) \text{ and } d_N(P_V, P_C) \leq d_N(P'_V, P_C)$$

Proof:

After splitting X_k into X_{k_1} and X_{k_2} in P'_V, we will have

$$|X_k \cap C_l| = |X_{k_1} \cap C_l| + |X_{k_2} \cap C_l|.$$

Therefore, $P_{kl} = P_{k_1 l} + P_{k_2 l}$. Now, the difference in the computation of $d(P_V, P_C)$ with respect to the computation of $d(P'_V, P_C)$ is that the terms:

$$- P_{kl} log_2 \frac{P_{kl}}{P_l} \quad (27)$$

and

$$- P_{kl} log_2 \frac{P_{kl}}{P_k} \quad (28)$$

intervening in the computation of $d(P_V, P_C)$, will be respectively substituted by:

$$- \left(P_{k_1 l} log_2 \frac{P_{k_1 l}}{P_l} + P_{k_2 l} log_2 \frac{P_{k_2 l}}{P_l} \right) \quad (29)$$

and

$$- \left(P_{k_1 l} log_2 \frac{P_{k_1 l}}{P_{k_1}} + P_{k_2 l} log_2 \frac{P_{k_2 l}}{P_{k_2}} \right) \quad (30)$$

in the computation of $d(P'_V, P_C)$.

Because X_k is split randomly into X_{k_1} and X_{k_2}, we have $P_{k_1 l}/P_{k_1} = P_{k_2 l}/P_{k_2} = P_{kl}/P_k$, so the terms (28) and (30) are equal. But (29) is greater than (27), because when $p = p_1 + p_2$ and $p, p_1, p_2 \in [0, 1]$ we have that $- log_2 p \leq - log_2 p_1$; and $- log_2 p \leq - log_2 p_2$. Therefore $d(P_V, P_C) \leq d(P'_V, P_C)$.

Finally, let us also prove that $d_N(P_V, P_C) \leq d_N(P'_V, P_C)$

Proof:

In this case, besides the replacement of (27) and (28) by (29) and (30) in the numerator, the term $- p_{k_l} log_2 p_{kl}$ intervening in the denominator is also replaced by $- (p_{k_1 l} log_2 p_{k_1 l} + p_{k_2 l} log_2 p_{k_2 l})$. We have then that the increase in the numerator is:

$$\delta_N = -\left(P_{k_1l}\, log_2 \frac{P_{k_1l}}{P_l} + P_{k_2l}\, log_2 \frac{P_{k_2l}}{P_l}\right) + P_{kl}\, log_2 \frac{P_{kl}}{P_l}$$

and the increase in the denominator is:

$$\delta_D = -\,(P_{k_1l}\, log_2\, P_{k_1l} + P_{k_2l}\, log_2\, P_{k_2l}) + P_{kl}\, log_2\, P_{kl}$$

Now, since $P_{kl} = P_{k_1l} + P_{k_2l}$ it is trivial to check that $\delta_N = \delta_D = \delta$.
Therefore, we finally have:

$$d_N(P'_V, P_C) = \frac{d(P_V, P_C) + \delta}{I(P_V, P_C) + \delta} \geq \frac{d(P_V, P_C)}{I(P_V, P_C)} = d_N(P_V, P_C)$$

because

$$\frac{d(P_V, P_C)}{I(P_V, P_C)} \leq 1$$

We have then proved that our distance does not favor attributes with large numbers of values.

5. Experimental results

Recent comparative studies of several selection measures for decision-tree induction (Mingers, 1989) show that Quinlan's Gain Ratio generates the smallest trees. We have compared our distance-based criterion with Quinlan's Gain Ratio using data of two medical domains (see Table 1) already used by other researchers (Clark & Niblett, 1987; Cestnik et al., 1987) to compare their inductive algorithms. In each domain we have taken different proportions (60%, 70%, 80%) of randomly selected examples for training and the remaining (40%, 30%, 20%) for testing. For each proportion we performed 6 runs. The average complexity and accuracy over the 6 runs of the induced trees is shown in Tables 2 and 3 for hepatitis and breast cancer respectively.

Our results support previously reported results (Breiman et al., 1984; Mingers, 1989) that the accuracy of the induced trees is not sensitive to the goodness of the attribute selection measure. However, our distance may generate smaller trees than Quinlan's Gain Ratio especially in the domain of hepatitis whose attributes have a large variability in the number of values, although the differences are not statistically significant.

Table 1. Characteristics of the data sets

Domain	No. of Classes	No. of Attributes	No. of Examples	Values/Attributes
Hepatitis	2	19	155	(9,2,2,2,2,2,2,2,2,2,2,2,7,7,7,7,1,0,2)
Breast cancer	2	10	288	(3,2,3,3,2,2,3,2,5,2)

Table 2. Results for the Hepatitis data

Prop.	No. of Leaves (Gain ratio, distance)	Accuracy (Gain ratio, distance)
60%	(19, 18)	(77.9, 77.0)
70%	(20, 18)	(78.6, 79.3)
80%	(24, 20)	(80.0, 80.0)

Table 3. Results for the Breast cancer data

Prop.	No. of Leaves (Gain ratio, distance)	Accuracy (Gain ratio, distance)
60%	(73, 71)	(68.3, 70.7)
70%	(79, 78)	(69.2, 70.6)
80%	(87, 87)	(69.1, 70.2)

6. Conclusions

The aim of this note was to introduce a distance between partitions as an attribute selection criterion to be used in ID3-like algorithms. We have also shown the relation between our distance and Quinlan's Gain criterion by reformulating Quinlan's Gain in terms of measures of information on partitions. Such a relationship provides an interesting interpretation of Quinlan's normalized Gain as a similarity relation, and this helps to clarify its meaning. Furthermore, we have formally shown that our distance does not favor attributes with larger ranges of values. Thus, we have a clean, non *ad hoc* measure that does as well (or slightly better) in its performance compared to the previously thought best measure (i.e., Quinlan's Gain Ratio used in conjunction with the original Gain measure). We intend to pursue this comparison further with more data sets. We also believe that our formal analysis provides the "proper" normalization for Quinlan's Gain.

Acknowledgments

I am grateful to Bojan Cestnik and Matiaz Gams from the "Jozef Stefan" Institute of Ljubljana (Yugoslavia) for providing the data for breast cancer and hepatitis and to J.J. Crespo for his assistance in performing the experiments.

Thanks also to Walter van de Velde for helping me to improve an earlier version of this paper, to Ivan Bratko for his comments and to Donald Michie for pointing out to me that the normalization of my distance was the "proper" normalization for Quinlan's gain. The comments of anonymous reviewers also improved the final version of this paper.

References

Bratko, I., & Kononenko, I. (1986). *Learning diagnostic rules from incomplete and noisy data.* Seminar on AI Methods in Statistics. London.

Breiman, I., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regressing trees.* Belmont, CA: Wadsworth International Group.

Cestnik, B., Kononenko, I., & Bratko, I. (1987). ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In I. Bratko & N. Lavrac (Ed.), *Progress in machine learning,* Sigma Press.

Clark, P., & Niblett, T. (1987). Induction in noisy domains. In I. Bratko, & N. Lavrac (Eds.), *Progress in machine learning,* Sigma Press.

Hart, A. (1984). Experience in the use of an inductive system in knowledge engineering. In M. Bramer (Ed.), *Research and developments in expert systems.* Cambridge University Press.

Kononenko, I., Bratko, I., & Roskar, E. (1984). *Experiments in automatic learning of medical diagnostic rules.* (Technical Report) Ljubljana, Yugoslavia: Jozef Stefan Institute.

López de Mántaras, R. (1977). *Autoapprentissage d'une partition: Application au classement itératif de données multidimensionelles.* Ph.D. thesis. Paul Sabatier University, Toulouse (France).

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

Quinlan, J.R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert systems in the microelectronic age.* Edinburg University Press.

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