When Does Overfitting Decrease Prediction Accuracy in Induced Decision Trees and Rule Sets?*

Cullen Schaffer
Departments of Computer Science and Statistics
Rutgers University
New Brunswick, NJ · 08903
201-932-3145 · schaffer@paul.rutgers.edu

Abstract

Researchers studying classification techniques based on induced decision trees and rule sets have found that the model which best fits training data is unlikely to yield optimal performance on fresh data. Such a model is typically *overfitted*, in the sense that it captures not only true regularities reflected in the training data, but also chance patterns which have no significance for classification and, in fact, reduce the model's predictive accuracy. Various simplification methods have been shown to help avoid overfitting in practice. Here, through detailed analysis of a paradigmatic example, I attempt to uncover the conditions under which these techniques work as expected. One auxilliary result of importance is identification of conditions under which overfitting does *not* decrease predictive accuracy and hence in which it would be a mistake to apply simplification techniques, if predictive accuracy is the key goal.

*The research reported here was supported by a grant from the Robert Wood Johnson Pharmaceutical Research Institute.

1 Fitting and Overfitting: The Simplest Case

A certain fictitious kind of snail may have either one horn or two. A randomly selected snail has one horn and is observed to prefer apple mash to banana mash. A second randomly selected snail also has one horn and shows the same preference. A third randomly selected snail, however, has two horns and prefers banana mash. A fourth snail is now selected and we are asked to predict its taste preference on the basis of the three we have just observed.

If the new snail has two horns, we are faced with a dilemma. On the one hand, two of three observed snails prefer apple mash, suggesting that we should predict this preference. On the other hand, in our limited experience, two-horned snails have always been observed to prefer banana mash, suggesting the opposite answer.

On the basis of the evidence, we would estimate the probability that a randomly selected snail will prefer apple mash as $\frac{2}{3}$ and the probability that a randomly selected two-horned snail will prefer banana mash as 1. Of these two estimates, the first is more reliable, since it is based on more data, while the second is more pertinent to our prediction problem. The difficulty of the problem, from one perspective, is in weighing these incommensurable advantages.

A second, more familiar perspective is that the difficulty lies in deciding which patterns in the data reflect true regularities in snail behavior and which are due solely to chance. In effect, we have been asked to build a model on the basis of data. The question is how to capture true regularities in this model, while excluding spurious ones—that is, how best to *fit* the model without *overfitting* it.

The problem of overfitting is well documented in reports on decision tree and rule induction approaches to classification tasks; experience in practical applications has shown that the tree or rule set which best fits training data is rarely the one that will perform best in classifying new cases. Research has generally focused, however, on practical methods for avoiding overfitting rather than on deepening our understanding of it.

This paper concentrates, instead, on the latter goal and attempts to provide insight into the conditions under which overfitting adversely affects prediction accuracy. In approaching these questions, I have begun by isolating what I believe is the smallest rule induction or decision tree problem for which overfitting is a concern. An analysis of the snail problem—the result

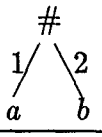
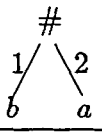
	Model	Prediction	Rule Set	Decision Tree
Simple Models	M_1	a	T $\Rightarrow a$	a
	M_2	b	T $\Rightarrow b$	b
Complex Models	M_3	a for 1, b for 2	$1 \Rightarrow a$ $2 \Rightarrow b$	
	M_4	b for 1, a for 2	$1 \Rightarrow b$ $2 \Rightarrow a$	

Table 1: Models for the Snail Problem

of this distillation—is the main focus in what follows. By looking at overfitting in the small, however, I will argue that we learn about dealing with the problem as it arises in practice.

2 A First Result

Given a series of observations, each consisting of a number of horns (1 or 2) and a taste preference (a or b), only four deterministic models of preference are possible. Table 1 lists these and shows that we may represent the models equally well as decision trees or rule sets.

In some cases, data may unambiguously support one of these four models. For example, the observations $\{<1,a>,<1,a>,<2,a>\}$ clearly favor M_1. In other cases, the evidence lends no special support to either of the complex models M_3 and M_4. Given the observations $\{<1,a>,<2,a>,<2,b>\}$ or $\{<1,a>,<1,a>,<1,b>\}$, for example, M_1 fits exactly as well as M_3. What remain are *equivocal* cases, exemplied by the case of the previous section, which lead to a choice between a simple hypothesis and a more tenuously supported, but better fitting complex hypothesis.

Employing this new term, I can state the concern of this paper precisely: I would like to ask about the conditions under which, in equivocal cases, choice of the best fitting complex model over the best fitting simple one leads to

a decrease in prediction performance. In particular, I will be investigating this question for training sets of size three, since these are the smallest which may be equivocal.

Let S_s be the strategy of choosing the best fitting simple model in equivocal cases and S_c be the strategy of choosing the best fitting complex model.[1] The question, then, is to determine conditions under which S_s is superior in predictive accuracy.

Let p_1 be the true probability of a one-horned snail preferring apple mash and let p_2 be the true probability of a two-horned snail preferring apple mash. Suppose, temporarily, that we have no prior knowledge about taste preferences in snails and, hence, none about the values of p_1 and p_2. We might model this state of ignorance by assuming, *a priori*, that any (p_1, p_2) pair is as likely as any other, that is, that the prior joint distribution for these parameters is uniform on the unit square.

Given this assumption, a straightforward Bayesian analysis allows us to calculate the expected prediction accuracy of models chosen by S_s and S_c for three-observation, equivocal training sets, averaged over the possible values of p_1 and p_2.[2] This analysis shows that S_c yields rules with a higher average prediction accuracy than S_s. That is, under the conditions described, overfitting should *not* be a concern—the model which best fits the training data is the one which is likely to do best in future prediction.

Given the effort researchers have expended in devising techniques to avoid overfitting in inducing classification models, this result is rather striking at first. The catch is that it depends on the assumed prior distribution for p_1 and p_2. This implicitly entails an assumption that p_1 and p_2, considered as random variables, are independent and hence it necessarily makes observations of one-horned snails irrelevant to predictions about two-horned ones. In essence, we have assumed *a priori* that we must adopt the complex strategy of estimating the majority class for the two groups separately.

This initial result illustrates how we may study the effect of overfitting on prediction accuracy and specify conditions under which a bias toward simplicity is or is not desirable. It also shows that the choice of a prior distri-

[1]Note that, while the simplicity of models M_1 and M_2 is purely syntactic and hence arbitrary, the simplicity of S_s relative to S_c is not. To fit a model, S_c estimates the majority class for each of two groups; S_s does so for just one.

[2]I also assume, throughout the paper, that the probability of selecting a one-horned snail is .5 both for the training and the test sets.

bution for p_1 and p_2 is critical in weighing S_s against S_c. The strongest lesson of this first result, however, is that the straightforward Bayesian approach of choosing a prior distribution and analyzing its implications may be dangerously unintuitive; assumptions entailed by a prior distribution are often less than obvious. Rather than burying such assumptions in a joint distribution, I have adopted a "disguised Bayesian" approach in the following which I believe is better suited to the goal of promoting intuitive understanding.

3 A Disguised Bayesian Analysis

Assume, then, that values for p_1 and p_2 are fixed, though unknown to us. Suppose, for example, that the values are, respectively, .6 and .8. In this case, the simple model M_1 is clearly optimal in prediction. It does not follow, however, that S_s is the optimal model-selection strategy under the assumed conditions. Sheer chance may lead to the observation sequence $\{<1,b>,<1,b>,<2,a>\}$, in which case S_s will select M_2 with accuracy .3 while S_c will select M_4 with accuracy .6. Of course, for the assumed values of p_1 and p_2, we would not expect this sort of problem to arise very often; S_s ought *normally* be superior. Formally, we may calculate the performance of models selected by the two strategies for each equivocal observation sequence and average these performance figures, weighting by the chance of the various observation sequences arising under the assumed values of p_1 and p_2. This calculation confirms our intuition; since it shows that models chosen by S_s have an average prediction accuracy of .573 while those chosen by S_c have an average prediction accuracy of .560.[3]

By making similar calculations for each possible pair of p_1 and p_2 values, we may construct the diagram shown in Figure 1. Boundary lines in the figure cover (p_1,p_2) pairs for which S_s and S_c select models with the same average prediction accuracy. In the upper left and lower right regions, S_c is superior; in the lower left and upper right, S_s is superior.

Although it omits important details—such as the *degree* to which prediction accuracies for the two strategies differ at a given (p_1,p_2) point—a diagram of this kind gives us a useful intuitive perspective on conditions under which the conventional bias toward simplicity is justified. Essentially, S_s

[3]Details of a more general calculation are described in Section 7.

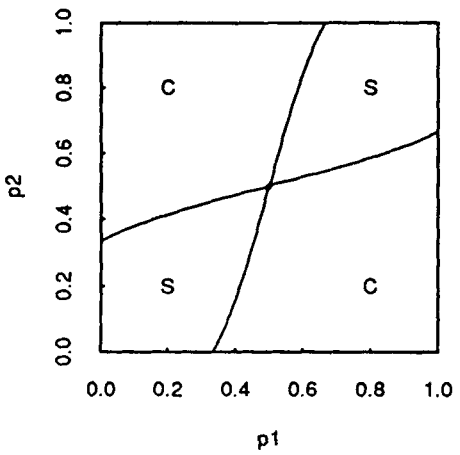


Figure 1: Comparing S_s and S_c

is preferable only for (p_1, p_2) pairs lying near the $p_1 = p_2$ line, though the acceptable distance grows as the parameters approach 0 or 1.

This result is easily explained. The more one- and two-horned snails are alike, the less is to be gained by treating them differently for purposes of prediction. At the same time, the risk incurred in considering the two types of snails separately remains constant. This risk is due to the fact that S_c relies on less data in determining predictions for each snail type than S_s; for the observation sequence $\{<1,a>,<1,a>,<2,b>\}$, it bases predictions for two-horned snails on just one observation.

The striking point about Figure 1 is not, then, that the S_s regions lie along the $p_1 = p_2$ line, but rather that they are so much smaller than those in which S_c is preferable. This would seem to suggest that a bias toward simplicity increases prediction accuracy only in exceptional circumstances. Since practical experience has shown exactly the opposite, however, we must ask what it is about practical applications that leads to a prevalance of parameter vectors falling in S_s regions. I will come back to this question in Section 5 after first discussing the effects of noise.

4 Noise

The analysis of the previous section implicitly assumes what Spangler, et al. (Spangler et al., 1988) have called "inconclusive data." If p_1 and p_2 are not taken from the set {0,1}, then the number of horns on a snail does not

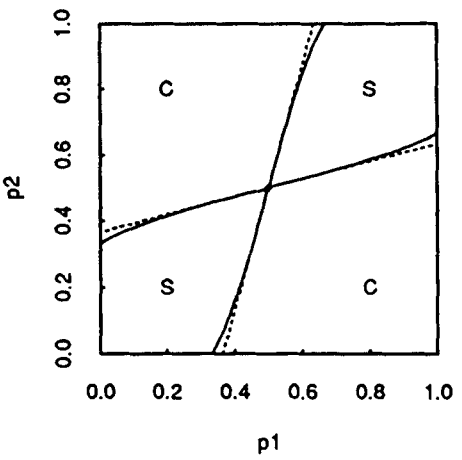


Figure 2: Comparing S_s and S_c, $n_c = 0, .8$

entirely determine its taste preferences. Presumably, as in many realistic application, unobserved attributes account for this apparent deviation from determinism.

For additional realism, we may consider the possibility of influences compromising the validity of some observations. If a particular one-horned snail happens to prefer apple mash, for example, *noise* of various kinds might cause us mistakenly to treat the observation as either $<1, b>$ or $<2, a>$ or even $<2, b>$.

Let n_d be the level of *description noise*, the probability that the true number of horns will be replaced by a random value equally likely to be 1 or 2. Let n_c be the level of *classification noise*, the probability that the true taste preference of a snail will be replaced by a random value equally likely to be a or b. For either of these kinds of noise, let a primed symbol denote the noise level in test (or performance) data, if this is different from the level in data on which a prediction model is based. Thus, n'_d denotes the level of description noise in the test data.

Unexamined intuition might suggest that the increase in uncertainty resulting from introduction of noise would tend to favor selection of S_s over S_c. Figure 2 shows, however, that this is only weakly so for the case of even extreme classification noise.[4]

This figure is, again, easy to explain. Adding classification noise has precisely the same effect as shifting the parameter pair (p_1, p_2) in a straight

[4]Details of calculations supporting figures in this section are given in Section 7.

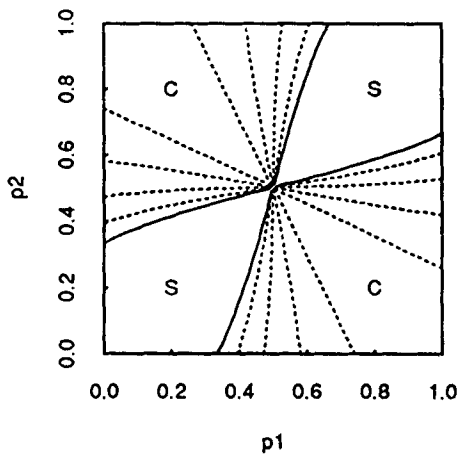


Figure 3: Comparing S_s and S_c, $n_d = 0, .2, .4, .6$ and $.8$

line toward the point (.5,.5). Examination of Figure 1 shows that very few points in the region where S_c is preferred can be moved into the region where S_s is preferred by such a translation. Hence, addition of classification noise expands the S_s region only by annexation of these points.

Description noise, however, obliterates the distinction between *description groups*—identifiable subsets of the population which may be treated differently by a classification model—and makes it less profitable to consider the groups separately. While classification noise shifts (p_1, p_2) in the direction of (.5,.5), description noise shifts it toward the $p_1 = p_2$ line. Examination of Figure 1 indicates that *all* points in the S_c region can eventually by moved into the S_s region by such a translation. Hence, the intuition that simple models are best for weak data is justified in this case. Figure 3 illustrates the effect graphically. Note that noise must be increased to a level of roughly $n_d = .4$ to make the size of the S_s and S_c regions comparable.

A last point to note about description noise is that n_d and n'_d each have a significant effect on the relative merit of S_s and S_c. Figure 4 shows first the effect of description noise in the test set (where it might most likely be expected to appear) and then the incremental effect of description noise in the training set.

5 Discussion

The results I have presented may be summarized as follows:

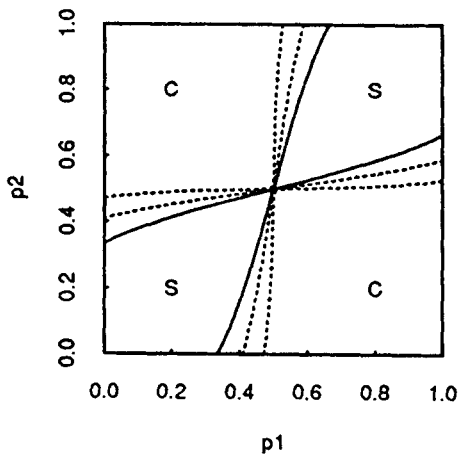


Figure 4: Comparing S_s and S_c, $n_d = n'_d = 0$; $n_d = 0, n'_d = .4$; $n_d = n'_d = .4$

- Whether overfitting ought to be a concern depends critically on the distribution of problems over the p_1-p_2 unit square.
- In the absence of description and classification noise, the portion of this square in which S_c is preferred is considerably larger than the one in which S_s is preferred.
- Classification noise has a nearly negligible effect on the relative merit of S_s and S_c.
- Description noise has a strong effect, however, even at moderate levels, and this is due in roughly equal measure to the effect of noise in the training and testing sets.

In response to the question posed in the title of this paper, then, we may tentatively generalize from the paradigmatic snail problem and answer that overfitting decreases prediction accuracy either (1) in the presence of significant description noise or (2) when classification patterns in description groups are not sharply distinguished in typical problems. The latter is just another way of stating the condition that available description attributes tend not to be very relevant to predicting classifications.

Given this summary, two questions are worth considering. First, since simplification strategies do generally help in practice, why is it that the conditions just stated normally hold?. Second, are there cases in which they do not hold and, hence, in which techniques for avoiding overfitting might actually yield inferior models?

Regarding the first question, description noise *is* certainly a common problem and, when present, may account for the fact that simplification techniques increase prediction accuracy. In many classification problems, however, description noise is insignificant. We may expect, for example, that a patient's vital signs will normally be measured correctly; the problem with using these to build a decision tree for diagnosis is that they do not entirely determine an illness.

If it pays to avoid overfitting in such cases, it must be because description attributes are not often highly relevant to classification. This may seem counter-intuitive at first, since we might expect analysts applying decision tree or rule induction techniques to choose attributes which they have good reason to believe *are* highly relevant. In fact, however, analysts often include a large number of description attributes, relying on the power of the induction technique to identify relevant ones. Also, perhaps more important, even relevant attributes are often highly interdependent, measuring similar qualities in a variety of ways. Hence, even if an attribute is highly relevant to classification taken by itself, it may be virtually valueless *after* other attributes have been taken into account. At the lower levels of a large decision tree, for example, the assumption that remaining attributes are unlikely to help might typically be justified.

These points lead to a natural answer to the second question posed above. In some cases, we may reasonably expect that description attributes are relevant and that description noise will not play a role. In these cases, some familiar and highly successful simplification techniques might be expected to lead to suboptimal classification procedures. For example, if we observe that two randomly selected dogs enjoy a certain food F, but a randomly selected person does not, it might well be reasonable to predict that people generally will not like F, despite the fact that two of three animals have been observed to enjoy it. Here, we are relying on our ability to distinguish reliably between dogs and people and on our common sense knowledge that an animal's species has a strong bearing on its taste preferences.

By contrast—to take a well-known example—Quinlan's (Quinlan, 1987) normally powerful technique of building a decision tree, converting it to a rule set and then simplifying this set would lead to the opposite prediction, since the observed difference between canine and human tastes is not statistically significant. The basic problem is that the statistical approach proceeds by assuming that attributes are irrelevant to classification unless proven other-

wise. This null hypothesis is precisely the second condition I have identified as conducive to the success of simplification techniques.

6 Related Work and Comments

In their seminal work on CART, Breiman, et al. (Breiman et al., 1984) analyze the problem of overfitting nicely, identifying the basic tradeoff between model support and potential accuracy. Their analysis concentrates on explaining why overfitting is a problem rather than on when this problem arises. Others, notably (Niblett, 1987) and (Weiss, 1987), have also suggested powerful approaches to dealing with the problem of overfitting in practice and attempted mathematical analyses to support their work. Like Breiman et al., these authors concentrate in both algorithms and mathematical analyses on handling the kinds of cases normally faced in practice—in which overfitting is a concern—rather than distinguishing these from the less likely cases in which methods for avoiding overfitting might actually decrease predictive accuracy.

My distinction between description and classification noise is drawn from Quinlan's analysis of the effect of noise on decision tree performance (Quinlan, 1986). Quinlan does not, however, address the question of how noise affects overfitting and strategies for avoiding it.

A good review of such strategies is given in (Quinlan, 1987). As this review makes clear, there are other reasons than prediction accuracy to prefer simplicity in induced classification procedures.

Needless to say, the tiny problem I have analyzed here assumes away a host of important subtleties and we must be careful to consider the extrapolations I have offered as suggestive rather than conclusive. Large data sets, large attribute sets, problems involving continuous or multi-valued attributes, prior knowledge of relations between attributes—all these affect the analysis drastically and may in some cases render it inapplicable. On balance, however, I think even this small problem deepens our understanding of the problem of overfitting. In a word, it suggests (1) that methods for dealing with overfitting work only because of the validity of certain implicit assumptions and (2) that we may benefit by attempting to make these explicit.

7 Appendix

This appendix explains how to calculate the prediction accuracy of S_s and S_c on the basis of given values for p_1, p_2, n_d, n'_d, n_c and n'_c.

If n_d is the probability that a description value will be replaced at random, then the chance that the observed value will be erroneous is $n_d/2$. Define $e_d = n_d/2$ and e'_d, e_c and e'_c analogously. Then the probabilities of making any of four basic observations are:

$$\begin{aligned}
P(<1,a>) &= [(1-e_d)(1-e_c)p_1 + (1-e_d)e_c(1-p_1) + \\
&\quad e_d(1-e_c)p_2 + e_d e_c(1-p_2)]/2 \\
P(<1,b>) &= [(1-e_d)(1-e_c)(1-p_1) + (1-e_d)e_c p_1 + \\
&\quad e_d(1-e_c)(1-p_2) + e_d e_c p_2]/2 \\
P(<2,a>) &= [(1-e_d)(1-e_c)p_2 + (1-e_d)e_c(1-p_2) + \\
&\quad e_d(1-e_c)p_1 + e_d e_c(1-p_1)]/2 \\
P(<2,b>) &= [(1-e_d)(1-e_c)(1-p_2) + (1-e_d)e_c p_2 + \\
&\quad e_d(1-e_c)(1-p_1) + e_d e_c p_1]/2
\end{aligned}$$

As noted earlier, $P(<1,\cdot>)$ is assumed to be .5.

If we consider the order of observations, there are 12 possible equivocal three-observation training sets. Strategies S_s and S_c ignore the order of observations, however, and we can simplify calculations by recognizing just four equivocal observation sets:

$$\begin{aligned}
O_1 &= \{<1,a>,<1,a>,<2,b>\} \\
O_2 &= \{<1,b>,<1,b>,<2,a>\} \\
O_3 &= \{<1,a>,<2,b>,<2,b>\} \\
O_4 &= \{<1,b>,<2,a>,<2,a>\}
\end{aligned}$$

Each of these may appear in three equally likely permutations. Hence, if we let

$$\begin{aligned}
k_1 &= P(<1,a>)^2 P(<2,b>) \\
k_2 &= P(<1,b>)^2 P(<2,a>) \\
k_3 &= P(<1,a>) P(<2,b>)^2 \\
k_4 &= P(<1,b>) P(<2,a>)^2
\end{aligned}$$

and

$$n = k_1 + k_2 + k_3 + k_4$$

then we have $P(O_i) = k_i/n$ for $i = 1, 2, 3, 4$. Note that these probabilities are implicitly conditioned on observation of an equivocal training set, since we are only comparing the accuracy of S_s and S_c under these circumstances.

Let $A(M_i)$ be the accuracy of M_i as measured by the probability of its predicting the observed class. Then

$$A(M_1) = (1 - e'_c)(p_1 + p_2)/2 + e'_c(1 - (p_1 + p_2)/2)$$
$$A(M_2) = e'_c(p_1 + p_2)/2 + (1 - e'_c)(1 - (p_1 + p_2)/2)$$

Since S_s chooses M_1 given O_1 or O_4 and M_2 given O_2 or O_3, we have

$$A(S_s) = [P(O_1) + P(O_4)]A(M_1) + [P(O_2) + P(O_3)]A(M_2)$$

where the accuracy of a strategy is taken to be the average accuracy of the models it selects. Likewise, we have

$$\begin{aligned}
A(M_3) = {} & [(1 - e'_d)(1 - e'_c)p_1 + (1 - e'_d)e'_c(1 - p_1) + \\
& e'_d(1 - e'_c)p_2 + e'_d e'_c(1 - p_2) + \\
& (1 - e'_d)(1 - e'_c)(1 - p_2) + (1 - e'_d)e'_c p_2 + \\
& e'_d(1 - e'_c)(1 - p_1) + e'_d e'_c p_1]/2 \\
A(M_4) = {} & [(1 - e'_d)(1 - e'_c)(1 - p_1) + (1 - e'_d)e'_c p_1 + \\
& e'_d(1 - e'_c)(1 - p_2) + e'_d e'_c p_2 + \\
& (1 - e'_d)(1 - e'_c)p_2 + (1 - e'_d)e'_c(1 - p_2) + \\
& e'_d(1 - e'_c)p_1 + e'_d e'_c(1 - p_1)]/2
\end{aligned}$$

and, since S_c chooses M_3 given O_1 or O_3 and M_4 given O_2 or O_4,

$$A(S_c) = [P(O_1) + P(O_3)]A(M_3) + [P(O_2) + P(O_4)]A(M_4)$$

As I have noted, these calculations yield the expected *observed* accuracy of S_s and S_c, counting a prediction model as successful when its predictions match the *apparent* class of new objects. To calculate the accuracy of these strategies in predicting the *true* class of new objects, simply let $n'_c = 0$.

8 Acknowledgements

Thanks to Tom Ellman, who suggested viewing noise as a model shift in the p_1-p_2 plane.

References

Breiman, L., Friedman, J., Olshen, R., Stone, C. *Classification and Regression Trees.* Wadsworth and Brooks, 1984.

Niblett, T. Constructing decision trees in noisy domains. In *Proceedings of the Second European Working Session on Learning*, pages 67–78. Sigma Press, Bled., Yugoslavia, 1987.

Spangler, S., Fayyad, U., Uthurusamy, R. Induction of decision trees from inconclusive data. In *Proceedings of the Fifth International Workshop on Machine Learning*, pages 146–150, 1988.

Quinlan, J. The effect of noise on concept learning. In Michalski, R., Carbonell, J., Mitchell, T. *Machine Learning: An Artificial Intelligence Approach*, volume 2, chapter 6. Morgan Kaufmann, 1986.

Quinlan, J. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27:221–234, 1987.

Weiss, S., Galen, R., Tadepalli, P. Optimizing the predictive value of diagnostic decision rules. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, pages 521–526, 1987.