

some amount of search should be preferred to an apparently superior rule found after more extensive search. This criterion leads to a method for curtailing search and we report results demonstrating the benefits of this strategy both for finding individual rules and for learning (complete theories). Finally, we offer limited evidence for the proposition that oversearching is orthogonal to overfitting.

2 Learning Individual Rules

This paper addresses the familiar propositional formalism in which each item belongs to one of its discrete classes and is specified by its values for a fixed collection of attributes [Quilan 1993]. The goal is to learn a classifier from a training set that predicts classes of unseen items. We concentrate on classifiers expressed as a sequence of rules of the form

if T_1 *and* T_2 *and* ... *and* T_u *then class* C_T

where a test T_j takes one of four forms: $A_j = t$ or $A_j \neq t$ for discrete attribute A_j , and value t ; or $A_j < v$ or $A_j > v$ for continuous attribute A_j and constant threshold t .

In the first experiment we focus on learning single rules following Webb [1993] in searching for one that minimizes the Laplace predicted error. Define the true error rate of a rule as the probability that an item that satisfies the rule's left-hand side does not belong to the class given by its right-hand side. If a rule such as the above is satisfied by n training items, c of which belong to classes other than the class C_x nominated by its right-hand side, the estimated error rate of the rule on unseen items is given by

$$\hat{\epsilon}(T \mid C) = \frac{c + A - 1}{n + k}$$

where k is again the number of classes.

To show the effects of increasing amounts of search, rules are found with beam search of width u varying exponentially from 1 to 512. For a given class C_r , the initial beam at level 1 consists of the u single tests that have the lowest Laplace error rate as above. At each subsequent level, with up to u conjuncts in the current beam, all ways of extending each conjunct with an additional test are considered and the best u of them retained for the next beam.

Notice that we can *prune* some combinations of tests without adding them to the beam. If a conjunct R matches n training items with e errors, adding further tests to R can only make it more specific and thereby decrease the number of items that it covers. A conjunct of the form R *and* S can thus do no better than match $n-t$ items with no errors. Unless $\hat{\epsilon}(TW, 0)$ is less than the Laplace error estimate of the best conjunct found so far, no descendant of R could ever improve on this best conjunct, allowing R to be discarded.

Search proceeds until the current beam is empty, whereupon the best conjunct found so far becomes the left-hand side of the rule for C_x .

We have carried out experiments on twelve real-world datasets from the UCI Repository that are described in

	Items	Classes	Attributes
breast cancer	286	2	4c 5d
house voting	435	2	16d
lymphography	148	4	18d
primary tumor	339	21	17d
auto insurance	205	6	14c 10d
chess endgame	551	2	39d
credit approval	690	2	6c 9d
glass	214	7	9c
hepatitis	155	2	6c 13d
Pima diabetes	768	2	8c
promoters	106	2	57d
soybean	683	19	35d

Table 1 Datasets used in the experiments

Table 1, the first four being the real-world domains studied by Webb. The size of each dataset, the number of classes, and the numbers of discrete (d) and continuous (c) attributes are shown. The following trial was repeated 500 times for each dataset:

Split the data randomly into 50% training and 50% test sets, making the class distributions as uniform as possible.
For beam widths $u = 1, 2, 4, \dots, 512$
For each class in turn
Identify the rule with lowest $\hat{\epsilon}$ value found during a beam search of width u .
Determine the rule's error rate on the test set.

Results of the experiments appear in Figure 1, where error rates are plotted against beam width. These error rates are weighted averages across the classes, the weights being the class relative frequencies in the training set. The dotted lines in each graph show the average $\hat{\epsilon}$ values of the rules selected, without exception $\hat{\epsilon}$ values decline with beam width as more extensively search discovers rules with lower predicted error rate. The solid lines, however, show the average true error rate of the rules as measured on the unseen test data. (The vertical bars show one standard error either side of the mean, the open circles flag the beam corresponding to the lowest true error rate, and the asterisks as explained in the next section.) As can be seen, the behavior of the true error rate is quite unlike that of the estimated rate $\hat{\epsilon}$. With some datasets such as the promoter domain, increasing search first lowers the true error rate, then causes it to rise, an example of the same non-monotonicity observed by Rymon [1993]. On other domains such as hepatitis, more extensive search is uniformly counter-productive. Only for the glass dataset does the true error rate of the selected rule decline non-monotonically with increased search.

To understand what is going on, we examine in more detail the chess endgame dataset, a particularly striking example of non-monotonicity. Separating results for the two classes (Figure 2), we can see that good rules for the majority class are found from the complete dataset with relatively small beam widths and thereafter in