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 rtport results demonstrating the benefits of this strategy both for finding individual rules and for learning (omplete theories Fmaly we offer limited evidence for the proposition that oversearthmg is orthogonal to overfitting

2 Learning Individual Rules

This paper addresses the familiar propositional formalism in which each item belongs to one of it discrete classes and is specified by its valess for a fixed collection of attributes [Quulan 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\ and T2 *and and* T_u *then class* C_T

where a test *T,* takes one of four forms A j = t or *A#* for disciete attribute 4, and value *v* and 4 j < f o r 4_,>/ for continuous attribute 4j and constant threshold *t*

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

$$£\{71\ C) = \frac{f + A - 1}{n + k}$$

where k is again the number of classes

To show the effects of increasing amounts of search rules art found with beam search of width u varving exponentially from 1 to 512 For a given class C_r the initial beam at level 1 consists of tht *w* single tests that have the lowest Laplace error rate as abovt A.t each subsequent level with up to u conjuncts in the current beam all wa\s of extending each conjunct with an additional test are considered and the bestn 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 An> conjunct of the form *R and S* can thus do no better than match *n-t* items with no errors Unless £ (T W , 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 tht 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 hrst four being tht real-world domains stuc led bv Webb The size of each datasct the number i classes and the numhers of discrete *(d)* and eontirn ous *(()* attributes are shown The following trial we repeated 500 times for each dataset

Split the data randomly into 50% trainnig and 50% test sets making the class distributions as uniform as possible

For beam widths u = 1 2 4 §12

For each class in turn

Identify the rule with lowest £ value found dunnq a beam search of width u

Determine the rule s error rate on the test set

Results of the se expe rime nts appear in Figure 1 in whre error rates are plotted against beam width These < ror rates are weighted averages across the classes tl weights being the class relative frequencies in The trail ing set Tht dotted lines in each graph show the ave age £ values of tht rults selected without cxceplmi £ values decline with beam width as more e xtcnsiy search discovers rules with lower predicted error rate The solid lines however, show the average true erre rate of the rules as measured on the unseen test dat. (The vertical bars show one standard error either sir of the mean, the open circles flag the beam corrcspone ing to the lowest true error rate and the asterisks ai explained in the next section) As can be seen the hi havior of tht true error rate is quite unlike that of tr estimated rate £ With some datasets such as the pre moter domain, increasing search first lowers the true e ror rate, then causes it to rise, an example of the san non-monotonicity observed bv Rymon [1993] On oth< domains such as hepatitis, more extensive search is un formly counter-productive Only for the glass datasi does the true error rate of the selected rule decline nea monotomcally with increased search

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