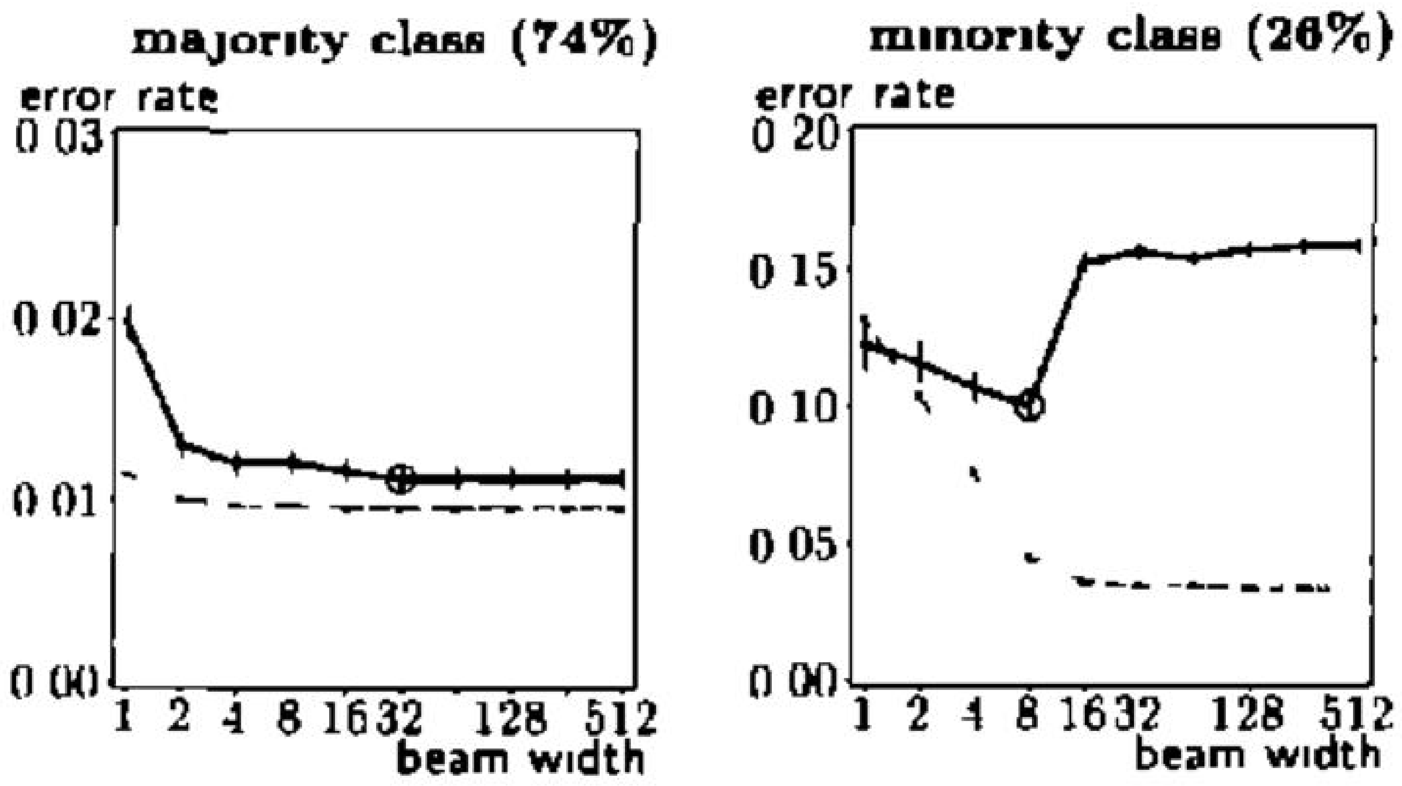

Figure 2 Chess endgame showing individual classes

Beam Width w	Items Covered e_u	n_u	Rules Examined h_u	Computed Estimate r_w
1	0	10	168	0 441
2	0	17	330	0 317
4	0	21	699	0 292
8	0	21	1265	0 311
16	0	23	2771	0 313
32	0	23	4758	0 329
64	0	23	7358	0 341
128	0	23	11768	0 354
256	0	23	17417	0 365
512	0	23	24902	0 375

Table 2 Selecting beam width

3 Selecting a Beam Width

Having e-stabhshed that extensive search can lead to less accurate rules we now discuss a method for limiting search

For the domains of Figure 1, the most accurate rule is often found with a beam width *w* greater than 1 (where *u=l* corresponds to greedy search) but less than 512 taken here as an approximation to exhaustive search Suppose now that a *layered search* were conducted b\ starting with *u<=* and doubling the beam width at each iteration Could we select the appropriate beam width so as to obtain the most accurate rule' This decision clearly cannot be made with reference to the £ \alue alone since this alwas decreases with further search

The following probabilistic argument was inspired b\ the famous Occam paper [Blumer Ehrenfeucht Haus-sltr and Warmuth, 1987] If the true error rate of a rule is r the probability that the rule will give no more than *e* e rrors m *n* trials is given bv

$$P(n \ e \ r) = \sum_{i=0}^{e} \binom{n}{i} r^i (1-r)^{n-i}$$

If then art *h* rules all having an error rate of r or inore the probability that an\ one of them will give e or less errors in *n* trials is at most $h \times P(n \ e \ r)$ whether or not the rules are independent

Now let h_u denote the number of rules examined during the search with beam width u, and let r„ satisfy

$$h_u \times P(n_u, e_u \ r_u) = 0 \, 5$$

If all these rules had error rate greater than or equal to TV there would be up to an even chance that one of them would give no more than r_u errors in n_u trials We use this value of r_u, as a gut estimate of t he accuracy of the best rule selected from the h_y candidates As *w* takes on the values 12 4 the corresponding values of h_u, n_v and e„, can be determined and the value of r_w computed We take the overall best rule to be that for which r_u is minimal

There are numerous over-simplifications m this argument For instance, it ignores the effect of beam selection at each level search for the rule with minimal £ value is guided bv the *C* values of partial rules, so that the $^k c_u$ errors in n„ trials is not a fair experiment Agam 'number of rules examined' is an imprecise concept manv putati\e rules cover no examples and some inks are pruned as described in See tion 2 For these expel iments h_u is taktn as the number of distinct altiibuu combinations considered during search on the basis that for each such combination there will be some test on every selected attribute that minimizes the inle s £ value

Table 2 illustrates the values for the positive class of the promoters dataset in one trial Greedy searc h finds a rule that covers 10 items without error Inecreading tin beam width to 2 causes a larger number of mles to be examined but vields a better rule covering 17 items still better rules are found at beam widths 4 and 1G In the latter case the number of rules examined mcieases (lie chance that the rule is a fiuke as reflected by its highe l r„ value The rule encountered at beam width =4 is consequently chosen as the overall best

We can now explain the asterisks in Figure 1 In each trial and for each class a best beam width is selected as above using only the training data The asterisk plots the average beam width selected and the average of the corresponding error rates on the unseen test data With the notable exceptions of the chess endgame and glass datasets, the average beam widths chosen are near the lowest points on the curves providing some empiric L1 support for the beam width selection strategy

4 Learning Complete Classifiers

The search for individual rules can be extended to learn complete classifiers using the standard covering method [Michalski 1980]

For each class C_x in turn
Mark all items of class C_T as uncovered
While uncovered items of class C_x remain
Find and retain the best rule
Mark as covered all class C_x items that satisfy the rule

The asterisk will not normally he on the solid curve because the beam width selected varies from class to class and from trial to trial