



Figure 2 Chess endgame showing individual classes

3 Selecting a Beam Width

Having established 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 $w=1$ corresponds to greedy search) but less than 512 taken here as an approximation to exhaustive search. Suppose now that a *layered search* were conducted by starting with $w=1$ 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 ϵ value alone since this always decreases with further search.

The following probabilistic argument was inspired by the famous Occam paper [Blumer, Ehrenfeucht, Hauslitz, and Warmuth, 1987]. If the true error rate of a rule is r , the probability that the rule will give no more than e errors in n trials is given by

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

If then start h rules all having an error rate of r or more, the probability that any one of them will give e or less errors in n trials is at most $h \times P(n \leq e)$ 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_u 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 r_u there would be up to an even chance that one of them would give no more than e_u errors in n_u trials. We use this value of r_u as a gut estimate of the accuracy of the best rule selected from the h_u candidates. As w takes on the values 12, 4, ..., the corresponding values of h_u , n_u and e_u can be determined and the value of r_u computed. We take the overall best rule to be that for which r_u is minimal.

There are numerous over-simplifications in this argument. For instance, it ignores the effect of beam selection at each level: search for the rule with minimal ϵ value is guided by the C values of partial rules, so that the

Beam Width w	Items Covered e_u	Rules Examined n_u	Rules Examined h_u	Computed Estimate r_u
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

h_u errors in n_u trials is not a fair experiment. Again 'number of rules examined' is an imprecise concept: many putative rules cover no examples and some links are pruned as described in Section 2. For these experiments h_u is taken as the number of distinct attribute combinations considered during search on the basis that for each such combination there will be some test on every selected attribute that minimizes the links ϵ value.

Table 2 illustrates the values for the positive class of the promoters dataset in one trial. Greedy search finds a rule that covers 10 items without error. Increasing the beam width to 2 causes a larger number of rules to be examined but yields a better rule covering 17 items. Still better rules are found at beam widths 4 and 16. In the latter case the number of rules examined increases (the chance that the rule is a fluke as reflected by its high r_u value). The rule encountered at beam width $w=4$ 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 null rates 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 usually the lowest points on the curves, providing some empirical 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 be on the solid curve because the beam width selected varies from class to class and from trial to trial.