Oversearching and Layered Search in Empirical Learning

J R Quinlan'
BasspT Department of Computer Science
Unnersity of Sydney
Svdnev 2006
Australia
quinlanOcs su oz au

R M Cameron-Jones
Department of Applied Computing
University of Tasmania
Launceston 7250
Australia
mcameronOleven appcomp utas edu au

Abstract

When learning classifiers, more extensive seaich for rules is shown to lead to lower prt- <liftiv(accuracy on many of the leal-world do-mains investigated Tihs counter-intuitive re suit is particularly lchvant to recent system the seaich methods that use nsk-fiee prun-ing to achieve the same outcome as exhaustive search We propose an iterated search method that commences with greedy search extending its scope at each aeration until a stopping cri-terion is satisfied This *layered* search is often found to produce theories that are more ac-curate than those obtained with either gree dy search or modcratrlv, extensive beam search

1 Introduction

Mitchell [1982] observes that the generalization implicit in learning from examples can beviewed as a search over the space of possible theories From this perspective most machine learning methods carry out a scries of local searches m the vicinity of tht current theorv selecting at each stop the most promising improvement Cover-ing algorithms like AQ [Michalski 1980] CN2 [Clark and Niblett 1989] and FOnl [Quinlan 1990] add new rules or Horn clauses to a developing theory divide-and-conquer methods such as c ARI [Biennan, Friedman, Olshen and Stone, 1984] and C4 5 [Quinlan, 1993] extend or re-vise a node of the current theory and selective mstance-based learners as exemplified by [Cameron-Jones 1992] add an item to the current set of retained instances

Theory spaces tend to be very large, so even these local searches must be constrained in the interests of efficiency Decision tree methods typically use greedy search (CART C4 5) or low-plv lookahead {CLS [Hunt, Marin and Stone, 1966]) while covering methods such as AQII and CN2 employ small-width beam search This limited search is guided by heuristics that are intended to identify simple theories consistent with the training set

This research was made possible by a grant from the Australian Research Council and assisted by research agreements with Digital Equipment Corporation

In stark contrast to this limited search Murphy and Pa/7am [1994] tackle the daunting task of generating ill such consistent decision trees In extensive experiments with four datasets they find that the smallest trees typ-ically have lower predictive accuracy than slightly larger trees, exhaustive search for the simplest consistent the-ories does not necessarily lead to improvement

Several investigators, notablv [Rvmon 1993 Schhm-mer 1993 Webb 1993], have recently developed brant h and-bound systematic search methods that have the same outcome as exhaustive search Again ths more extensive search has not led to the discovery of markedly better theories Rvmon reports non-monotonic improve-ment using three artificial datasets Webb describes opus, a system that resembles r N2 Both are covering algorithms that repeatedly look for a rule with minimal Laplace predicted error (discussed in Section 2) Despite the fact that OPI s effectivelv explores all rules whereas e N2 uses limited beam search the latter finds more pre-dictive theories on four of the five datasets studied

We believe that these rather discouraging results can be explained by noting that, for anv collection of train-ing data there are 'fluke[1] theories that fil the data well (according to whatever criterion is employed) but haye low predictive accuracy When a very large nuinbei of hypotheses is explored the probability of encountering sueh a fluke increases Since systematic search has the same outcome as exhaustive search it will always find such a fluke if one exists On the other hand heuristic search explores only a vanishingh small propoition of the space of theories and so is less likely 1o encounler a fluke It is commonly held that the construction of the ones that are more complex than can be justified by the data leads to poor predictive performance [Breiman *11 al* 1984, but see also Schaffer 1993] *Overfitting* refers to the construction of a theory tailored to the data that has high (but misleading) apparent accuracy By analogy we use the term *oversearching* to describe the discovery by extensive search of a theorv that is not necessarily oycr-complex but whose apparent accuracy is also misleading

In this paper we present empiric al evidence for the oversearching phenomenon and piopose a partial rem-ed> First, exploring larger numbers of potential theo-ries consistently leads to selection of better theories in only one of twelve domains investigated We deyelop a simple criterion for deciding whether a rule found after 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 Aj=t or *A#* for disciete attribute 4, and value *v* and 4j<for 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 £(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 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

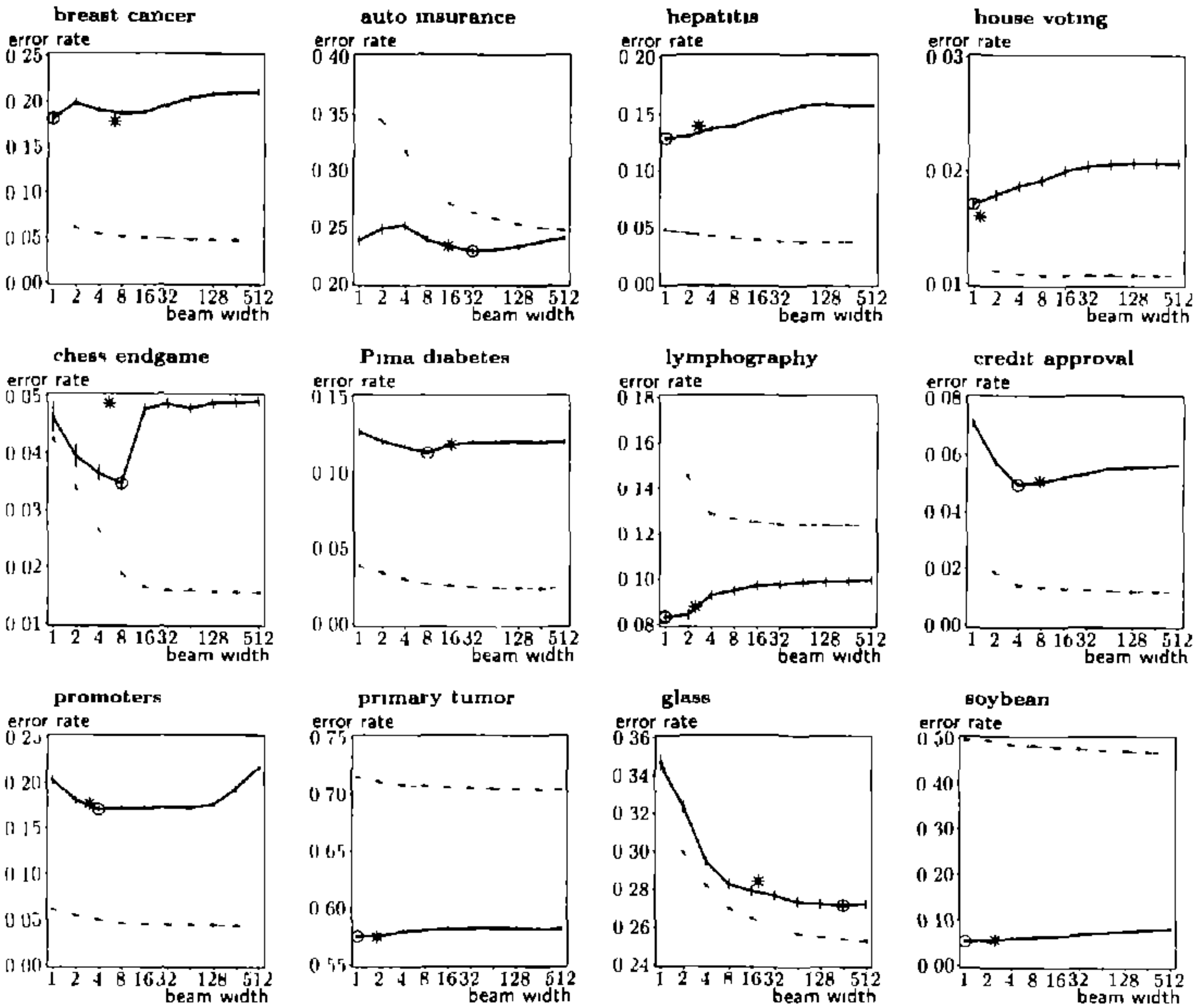

Figure 1 Effects of varying beam width

provenent is slight The I shape of the curve is due to th(mnioritv class fur wihch a marked change occurs at beam width 1G

In one typical tral search at beam width 8 finds a conjunction of three tests *(R1)* that is satisfied h\ 18 items of the minority class and none of the other class Further specialization of *Ri* can only decrease its cover and hence its £ value Howe\er there is also a conjunction of five tests *(R_2)* that covers 32 items of the minority class and *7* items of the other class Now, in order to discover a rule with left-hand side T1 *and* T_2 *and* *and* T_n, the beam at level *i* must contain at least one conjunction oft of these tests for all values of z from 1 to *n-l* Conjunct R_2 is difficult to find because no single test or pair of tests has a low £ value For this trial the £ value of the best single test ranks sixth among all single tests so *R2* is eliminated unless the beam width is at least G The best combination of two of the fi\e tests has an £ \alut that ranks thirteenth among all two-test combinations so the beam width needs to be al least 13 if *R2* is not to be eliminated at the second le\el of the beam search Once it is found however the large number of attnbules in this domain allows *R2* to be refined b\ the addition of seven further tests giving a rule R_3 that covers 30 nems without error In terms of the £ measure, R_3 *has* a lower predicted error than *R* and so is preferred

When evaluated on the test data howevei the com plex rule *Ra* misclassifies five items of the 31 that it matches, approximately the same error rate as the conjunct *R2* from which it was derived On the other hand, the rule *R1* is more accurate, misclassifving one of the thirteen items that it covers Increasing the beam width from 8 to 16 allows the "fluke' Ra to be discovered with a consequent increase in the error rate

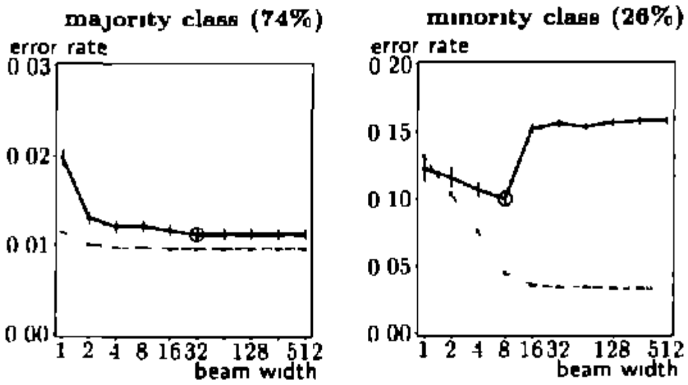

Figure 2 Chess endgame showing individual classes

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 Haussltr 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 \\ each trial and for each class a best beam width is selected as above using only the training data The astensk null rates the average beam width selec te d and the neiage of the corresponding error rates on the unseen test data ' With the notable exceptions of the chess endgame md glass datasets, the average beam widths chosen arc iu ai the lowest points on the curves, piovidmg some empiric L1 support for the beam width selec tion strategv

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

	Error Rate (%)			Number of Rules			Theory Size			Time (secs)		
	GS	LS	ES	GS	LS	ES	GS	LS	ES	GS	LS	ES
breast cancer	28 8	28 8	**29 1**	43 0	29 4	26 0	132 9	106 4	101 2	0 1	1 5	13 4
hoiiiie voting	5 7	5 6	5 7	14 3	10 9	10 1	37 5	31 7	30 3	0 1	0 4	11 1
lymphography	**22 1**	18 9	19 0	14 4	10 4	9 5	33 6	30 1	30 1	0 0	0 2	6 5
primary tumor	58 3	58 5	58 3	59 8	53 3	45 9	269 5	256 0	234 6	0 2	2 0	54 4
auto insurance	31 4	31 1	31 4	33 4	18 7	14 2	70 1	57 5	58 6	0 2	2 7	25 8
chess. endgame	**10 7**	10 3	**10 4**	44 0	28 9	27 3	130 7	112 2	113 7	0 3	4 7	150 9
(rrdiL approval	**16 7**	16 4	16 4	58 5	31 7	25 0	161 9	120 9	118 1	0 4	10 2	64 6
glass	**36 2**	34 1	**33 2**	27 3	18 1	15 6	74 2	59 8	56 5	0 1	1 1	8 0
hepatitis	**18 1**	19 1	**20 0**	14 3	10 4	9 4	29 7	27 0	27 9	0 1	0 3	1 9
Pima diabetes	**25 9**	26 9	**27 2**	96 3	50 1	44 3	301 2	207 4	208 7	0 8	14 8	34 1
promoters	**27 4**	24 6	**28 8**	8 3	5 5	4 1	16 4	13 5	14 1	0 0	0 2	4 2
soybean	**11 7**	12 4	**13 0**	39 4	35 9	29 5	112 4	108 9	106 6	0 4	2 4	67 3
Ratio to LS	*1 023*	*1 000*	*1 024*	*1 486*	*1 000*	*0 857*	*1 197*	*1 000*	*0 987*	*0 10*	*1 00*	*17 40*

Table 3 Results with greedv (GS) layered (LS) and extensive (ES) search

When (leterunning the best rule above onl\ uncovered items. of class C_T and all items of other classes are considtrcd Whereas Wlrbh [1993] finds the rule with the guaranteed lowest *C* value at each iteration, we use the best ruk encountered I)\ three kinds of h(unstir seaich

GS Greed\ search with beam width $u = 1$

LS Layered seanh witli beam widths u = 1 , *2 4,* 8 and so on ro a maximum of 512 For (cach beam width the rule with lowest *L* value entountcrcd during searrh is if tamed and its i„ value determined the DVERALL best rail best rule bemg the ond of these with lowest r„ Th(layered se-inh is terminated whenever two successive valuesof in fail lo improve on the best valu(of „ found so far

ES Fxtuisive seairh Willi fixed beam width *w=512* again taken to appioxnnate exhaustive search

An unsern item is classified by the ruleset b\ finding the rule with lowest £ value that matches it then assigning the item to the class speufied in that ruk s right-hand side item that satisfies no rule is assigned the most frequent class observed in the training set

The experimental design was similar to that described m Secttion 2 for each dataset 500 trials were conducted splitting the data into stratified equal-sized training and test sets Three classifiers were constructed from the training set using greedy (GS) layered (LS) and extensive (ES} search, respectively and each clabsiher evaluated on the test set Results averaged over the 500 repetitions appear in Table 3 A. simple indicator of theorv complexity is provided by *theory size,* the total number of tests in all rules Times are for a DEC AXP 3000/800 workstation

Those error rates for GS and ES shown in bold face are significantly[2] different from LS Layered search is significantly better than greexly search in five domains and worse in three When compared with extensive search layered search is significantly better in six domains and worse in only one Over the 6000 trials, LS is better than GS in 2822 trials and worse in 2534, while it is better than ES in 2927 trials and worse in 2438 both results are significant at better than p=0 0001

The ratio *to LS* figures in the final row g ve an overview across the twelve domains each is the average ratio of a result to that for layered search For these datasets, the theories found using LS have less than 98% of the error of those produced b\ either greedy or extensive search LS requires 10 times as much computation as GS but the absolute difference is small since the latter is so economical Extensive search (where u is hxed at 512) is 170 times slower than greedy search and 17 times slower than layered search even though the latter requires repeated search with increasing beam widths

[2]Two-tailed sign test p=0 05

5 Theory Complexity and Search

Discussion of the chess endgame example in Section *2* might suggest that this problem is just another instance of overfittmg extensive search is leading to the construction of elaborate rules Existing mechanisms for ovrrfitting avoidance such as Rissanen s Minimum DP scnptiom Principle [Quinlnlan and Rivest, 1989 Cameron Jones, 1992], might thus be sufficient to prevent the choice of rules with low predictive accuracy We offer two arguments against this hypothesis

\s can be seen in Table 3 ranking the search methods by the complexity of the theory produced does not correlate well with the accuracy of the theories Although ES often finds more complex individual rules this complexity is counterbalanced by their increased coverage Extensive search results in complete theories that arc simpler than those found by layered search and much simpler (20%) than those produced with greedy search Yet on average the FS theories are less accurate than their LS counterparts and have similar accuracy to the GS theories

The second is empirical based on preliminary e\ penments that assess the impact of oversearchmg on instances-based learning For these trials a classifier consists of a subset of the training items, with an unseen item assigned to the class of the most similar retained item All classifiers for a domain are constrained to ronsist of exactly the same number m of retained items, so that all theories have identical complexity Beam searches of various widths are again carried out, this time to find the m items that give the lowest classification error on the training set Results with the same twelve datasets are reminiscent of Figure 1 increased search leads to better and better sets of retained items as assessed on the training data, but the classifier s performance on unseen test data exhibits either a continuous decline or a I -shaped curve in six of the twelve domains

6 Conclusion

This paper provides further evidence that more search does not necessarily result in better learned theories In most of the domains studied litre expanding search leads eventually to a decline in predictive accuracy as idiosvncrasies of the training set are uncovered and exploited Thus phenomenon of oversearching has also been observed in other domains and indexe d with at least one other heuristic criterion "*

For the twelve datasets reported here an iterative lavered search with beam width limited bv a probabilistic criterion r_u was found to have better overall performance than either greedy or extensive search Even so, the argument that underpins the derivation of the r_u value and thereby selection of the best" beam width, is simplistic and we arc confident that a better criterion can be developed

We believe that oversearching cannot be controlled bv complexity-based mechanisms such as the MDL principle, the disadvantages of oversearching seem to be somehow orthogonal to problems of overfitting MDL is nghtlv popular because it provides a well-justified framework for mapping apparent accuracv and theory com pltxitv into a uniform measure based on coding length Ideallv we would like to see oversearching dealt with in a similarly clean manner by the development of a single metric that embodies all three factors accuracy, theory complexity and extent of search

Acknowledgements

We are grateful to Pat Langlev for his detailed and helpful comments on a draft of this paper, and to the anonvmous reviewers who suggested both improvements and areas for further research The breast cancer lymphography and primary tumor datast ts were provided by the Ljubljana Oncology Institute Slovenia Thanks to the LCI Repository and to its maintainers, Patrick Murphy and David Aha, for providing access to the datasets used here

[3] In place of the Laplace estimate we have also tried a confidence limit function *UCF* [Quinlan, 1993, page 41] This function turns out to be even more susceptible to coincidences in the training data, a majority of the domains discussed here show a monotonic decrease in predictive accuracv with increased beam width

References

[Blumer *et al*, 1987] Anselm Blumer Andrzej Ehrenfeucht, David Haussler, and Manfred Warmuth Occam s razor *Information Processing Letters 24* 377 380

[Breiman *et al*, 1984] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone *Classification and Regression Trees* Belmont Wadsworth

[Cameron-Jones, 1992] R Michael Cameron-Jones Minimum description length mstanc e-based learning *Proceedings Fifth Australian Joint Conferinct on Artificial Intelligence* Hobart Singapore World Scien tmc 368-373

[Claik and Niblett 1989] Peter Clark and Tim Nibhtt The (N2 induction algorithm *Machine Ltaming 1* 261-284

[Hunt *et al* 1966] Fail Hunt Tanet Mann and Philip Stone *Experiments in Induction* New York Ac deimc Press

[Michalski, 1980] Ryszard Michalski Pattern recognition as rule-guided inductive inference *IEEE Transactions on Pattern Analysis and Machine Inttelligenn S* 349-361

[Mitchell 1982] Tom Mitchell Generalization as search *Artificial Intelligence 18* 203-226

[Murphy and Pazzani 1994] Patrick Murphv and Michael Pazzani Exploring the decision forest an empirical investigation of Occam s razor in deeision tree induction *Journal of Artificial Intelligence Research I* 257-275

[Quinlan 1990] J Ross Quinlan Learning logical defi mtions from relations *Machine Lerinng 5* 239-2GG

[Quinlan 1993] I Ross Quinlan *C4 5 Pioqrtims for Machine Learning* San Mateo Morgan kaufmann

[Quinlan and Rivest 1989] T Ross Quinlan and Ronald Rivest Inferring decision trees using the Minimum Descnpl ion Length principle *Information and Computatum 80* 227-248

[Rymon 1993] Ron Rvmon An SE-tree based char aetcnzation of the induction problem *Proceedings Tenth International Conferenct on Machine Learning,* Amhcrst San Mateo Moigan Kaufmann, 268-275

[Sehaffer 1993] Cullen Schaffer Overfitting avoidantce < as bias *Machine Learning 10* 153-178

[Schhmmer, 1993] Jeffrey Schhmmer Efficiently indue ing determinations a complete and systematic search algorithm that uses optimal pruning *Proceedings Tenth International Conference on Machine Learning,* Amherst San Mateo Morgan Kaufmann 284-290

[Webb, 1993] Geoffrey Webb Systematic search for categorical attribute-value data-driven machine learning *Proceedings Sixth Australian Joint Conference on Artificial Intelligence,* Melbourne Singapore World Scientific, 342-347