Flexibly Exploiting Prior Knowledge in Empirical Learning

Julio Ortega
Computer Science Department
Vanderbilt University
Nashville Tennessee 37235
U S A
juho@vu&e \anderbilt edu

Doug Fisher
Computer Science Department
Vanderbilt University
Nashville, Tennessee 37235
II S \
dfisher@vuse vanderbilt edu

Abstract

This paper presents a method to incorporate knowledge from possibly imperfect models and domain theories into inductive learning of decision trees for classification The approach assumes that a model or domain theory reflects useful prior knowledge of th< task Thus the default bias should accept the model s predictions as accurate even in the face of somewhat contradictory data which may be unrepresenlative or noisy However our approach allows the svslem to abandon the model or domain theorv, or portions thereof in the fact of sufficientlv contradictory data In particular we use C4 5 to induce decision trees from data that ha\t heen augmented b\ model or domain-theory-denvcd features' We weakly bias the svslem to select model-derived features dur ing decision tree induction but this preference is not dogmatically applied Our experiments vary imperfection in a model the representa tiveness of data and the veracitv with which modfl-demed feature are preferred

1 Introduction

When human expertise is nonexistent or very weak relative to a particular domain/task and when data is plentiful machine induction from data mav be the only reasonable approach to task automation In contrast, when expertise is strong, then encoding the expert s model or domain theory via traditional knowledge acquisition strategies ma> be the best approach In fact, this human expertise may stem from induction over a much larger data sample than is available at the time task automation is undertaken

In many cases, however, conditions are indeterminate as to whether sole reliance on machine induction or human expertise is most appropriate human expertise may not be 'perfect and/or data may not be as plentiful as desired in cases where some data is available and human expertise is less than perfect an advantageous strategy may be to exploit both in an appropriate way

There is a growing body of work that combines model-based or domain-theory knowledge with empirical learning from data Clark and Matwin [1993] assume that an analyst-specified model mediates empirical learning - the rules derived from a machine-induction system are ace3pted as long as they do not contradict the biases found in the model Evans and Fisher [1994] employ a similar strategy - a human analyst may specify weak rules (e g when printing-plant humidity is low, a certain kind of printing error known as banding is more likely, to occur) If inductively-derived rules indicate an opposite trend then the learning system s default strategy is to reject the rule derived through induction In the case of both these approaches tht model or domain theory is deemed correct in its characterization of the domain task though it may not be a very deep characterization Inductive learning is used to flesh out rules that are con sistent with the model (e g by selecting the particular numeric thresholds that distinguish high from moderate and low) or discovering rules relevant to a part of the domain space that are not addressed by the model or weak domain theorv at all

In the approaches above if the data contradicts the model then the implicit assumption is that the data are noisy or unrepresentative drawn from a verv small subspace of the data Other approaches known is *theory revision methods* [Ourston l991] [Towell *et al* 1990] may give more credence to the data In these systems contradictions result in a revision of the domain theory to bring it in line with the data Drastal el al [1980] Rendell and Seshu [1990] and Ortega [1994] suggest an alternative strategy that loosely couples empirical learning and model-based reasoning tht data is augmented by features that are actually intermediate terms of the domain theorv and which are deemed true of a datum by deductive application of the domain theorv Induction is then performed over this augmented data set If domain-theory-derived features are included in rules denvea inductively, then this suggests a rough consistency between the model and data model features m iv be viewed as somewhat better predictors than raw features because noise is mitigated If model features are not referenced m a resultant classifier ihis may speak to imperfections in the model and/or this behavior mav stem from an unrepresentative data sample In bolh cases model-derived features ma\ not look as informative as raw' features relative to the *available* data

This paper describes a strategy that augments data with domain-theory derived 'features but unlike previ ous work we bias an adaptation of C4 5[Quinlan 1993] to select domain-theory based features even when this conflicts somewhat with C4 5 s original bias to select the most informative feature as computed over the data The intent is to guard against the possibility of unrepresentative data However the domain-theory preference bias mav be overridden if C4 5 s original bias is sufficient opposed to the domain-theorv preference bias The intent here is to acknowledge that there mav be some imperfections in the domain theorv Our evpenments van imperfection in a model the representativeness of data and the the veracitv with which modelderned features are preferred

2 Implementing a Flexible Domain-Theory Preference Bias

The approach described in this paper was motivated by our attempts to inductively build classifiers of faults of the Reaction Control System (RC S) of the Space Shuttle A mixed qualitative/quantitative model for fault pr< diction was available [Robinson 1993] as well as simulalted data representing svstem faults and normal behavior For each available datum the model was used to pre diet the fault This prediction was added as a featurr to the datum as were various intermediate computations made b\ the model for the data point The data points augmented in this way were then given to *CA ')* which constructed a classifier that predicted either a svstem s fault or normal operation If the model were perfect then we would expect that C4 5 would build a tret that only tested the model-based final prediction Such a tree would indicate that if a new datum w< re encountered (represented bv readings of various pressures and temperatures and other obsrrvables) then one should simply simulate the model on this datum and use the model-based final prediction In the cas« of certain imperfections a decision tree that tested vinous raw features as well as various model-based features might be constructed

To our initial surprise C4 5 consistently constructed trees that never or rarely referenced any model-based features Rather than taking this as evidence of significant model imperfection or that the model added little or no information, above and beyond that implieit in the raw features a NASA analvst familiar with this application indicated that the simulated data used for training was unrepresentative or skewed - it represented a verv small subspace of the RCS description space

This work motivated an approach that weakly biases our adaptation of C4 5 to select model-based features In particular for purposes of this paper we assume a propositional domain theory used for classification that is acyclic and directed from the observable propositions to a final classification A partial description of the perfect domain theory for the audiology domain used in our experiments is shown in Figure 1 as a tree The domain theory is a set of rules, each one consisting of a set of conditions together with the classification predicted b\ the rule In Figure 1 the antecedents of a rule are listed at the leaves of the tree Each condition is an attribute-value pair (e g Air=pTofound) There may be several rules thai predicL a particular claSMhcal ion as illustrated bv thf several possibh rules hading to each classification (e g OI ITIS MEDIA) in Figure I

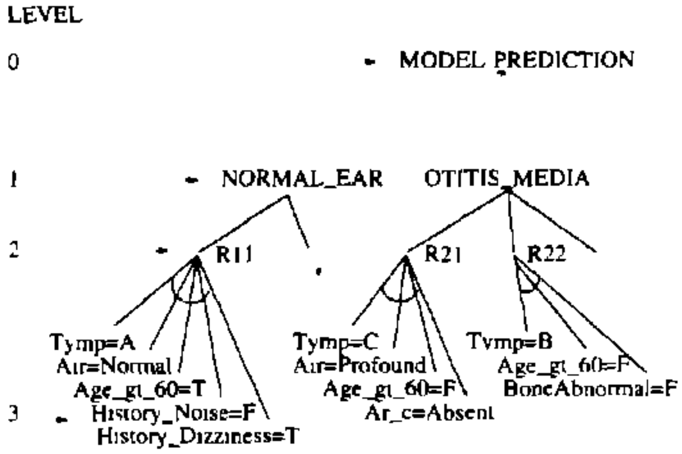


Figure 1 Levels in Audiology Theorv

We characterize model feitures extracted from a the one of this type according to their distance in the th<orv hierarchy to the final model prediction

- Level 0 Model prediction feature We generate d single model prediction feature Its value is the fin^l prediction mad' bv the rule interpreter (using the rules of the given theorv)
- Level 1 Int< rrnediate concept feature* One of these features is generated for each possible ela.s sification in the theory Each inte rmedial< concept feature corresponds to a logical OR of the rules that predict particular classifications
- Level 2 Rule features Features of this. tvpe are generad for each rule in the theory A rule feature follows from the logical AND of its antecedents
- Level 3 Raw Features Each rul* anti cedent is t binary test as to whether an attribute takes a par ticular valuf on an example Raw features whether rule antecedents or not are observable and are mi tiallv the exclusive means of representing data

To bias C1 ʳ> towards model features clost r in the hierarchy to the final model prediction, we order features according to their level number from the model predic tion feature through niternn tliate concept fiatures to rule featurrs, and raw features At each step during in duction our variation of (4 5 chooses a feature of smallest level number unless a statistically-significant better feature (in terms of C4 5 s information score) of larger level number is found Hence, C4 5 will choose the model prediction feature unless sufficient evidence is present in the data to refute this choice

Thus we bias our inductive algorithm toward the model prediction feature and other features closer to it (of small level number) In a situation where we have a reasonably accurate model, and the available data is unrepresentative we expect our model-biased method to work better than a default strategy of choosing the feature of highest information, value according to the available data (eg as in the standard C4 '>) Nonetheless, if the data sufficiently contradicts the model the model-bias can be abandoned and should we choose the model can be revised acrordinglv

3 Using Domain-Theory Bias with Hypothesis Testing

The major difference between the original and our variation of (A 5 is the manner in which a feature is selected for each nod*, of a decision tree C4 5 selects the feature with the highest information value according to th* *information gam ratio* measure Rather than selecting I lie fealure with the highest informal ion value outright our variation of C \ ') requires that this value be statistical^ significantly higher than the information value uf all features preceding it in a feature prefer'nee ranking like that described in the previous section Put in another wav we select the highest feature in a preference ranking that has an information score not significantly worse Ihan an\ feature lower in the preference ranking

The above procedure is implemented b\ the function SelertFature(\ *udt*) shown m Figure 2 where *Fp* is the feature preference ranking ' *I)* is the set of triming data associated with the current node *uifo[Fj 0)* is the vahii of (4 5 s information measure for feature r_j when evaluated on the set of dala *D* and *Fj* is the hsl of the features sorted in desce nding order according to (his measure SelectFeature(\ *ode)* initiallv chooses th^-feature v\ith highest information vilu^ (I < first feature m *F \)* However this feature is not accepted unless it> information value is *significantly higher* than all features of higher preference according to the *Fp* ranking If so the candidate fealure is stl<et«d Otherwise the high* r preference feature found bee omes the new candi date 1 be procedure is repeated unlil a significant dif fereiin is found or the F_1 list is < vhausted

Tht rt is also a minor difference between the cla-ssifi cation procedure of our system and the standard C4 5 algorithm for the situations where thert is insufficient data to select a lest for a particular node of the tree As a purelj data driven s\stem the best C4 5 can do is to predict the most common class present in the current node Instead since we assume our mode 1 is better Ihan no information we use the prediction of our prior model

The critical component of (he function SelectFeature is the SignificantlyBetter(/,-ııı,1f J_{yrj} *D)* function shown in Figure J This function returns true if the information value of feature $f_{ran(}t$ is estimated to be significantly higher than that of $f_{pr}ef$ according to a given level of statistical significance *^igLevel* I Ins is done by testing the null hypothesis that the difference between the information values of $f_{ra}nd$ and $f_{p}rtj$ is zero If (his null hypothesis can be rejected with 1 — *SigLtvel* confidence SigmficantlyBetter concludes that $f_{can}d$ *is* significantly better than $f_{p}rj$

[1]In the current implementation the ranking is a total or denng features are sorted in ascending according to level number The ranking of features within a level number is arbitrary

```
Given prior preference list F_I = f_1 f_2   f_F

Function SelectFeature( Node)

Set D to set of observations in Node
Create list of features F_I = f_j1 f_j2   f_jF_I
     sorted in descending order according to value
     of info(f_j D) eliminating any feature of null
     information value In the case of nominal
     features this precludes the consideration of a
     feature used previously in the same path
Set f_cand = f_j1
While no significant difference has been found
and there remain features to consider in F_I
     Set f_pref to the first feature in F_I after
          f_cand that precedes f_cand in F_F
          Eliminate all other features
          between f_cand and f_pref in F_I
          from consideration
     If Not(SignificantlyBetter(f_cand f_pref D))
          Set f_cand = f_pref
     EndIf
EndWhile
Return(f_cand)
```

Figure 2 Function SelectFeature

If the. form of the probability distribution associated with C 1 ') s information measure is known and its parameters can be calculated then traditional statistical thtorv can b(u&ed 1o test significance This could be done for Ihe information gain measure since Musick el al [Musick *e1 ai* 1993Q] prove that this measure is nor nnIlv distributed and provide exphent formulas for the parameters of this distribution However the form of the distribution for the default measure used in \ 4 5 mfoi mation gain ratio is not known Fortunately *Bootstrap Wethodt,* [Lfron and Giong 1983] allow for estimates of significance levels of arbitrary statistics when the form and parameters of the underlving distribution art not known [Moreen 1989] I Ins is the method implemented in the funrtion SigiuflcantlyBetter

In Lfron s Bootstrap methods an unknown complete population *P* is estimated b\ repeated uniform subsampling with replacement from an available sample *D* of *P* From ... obtain a. set of bootstrap subsamples $P_B = \{D_1 \ D_{N_B}\}$ where \Q IS a prespicihed number of subsarnples Each D_i (with $1 \leq i \leq N_B$) is very likely to contain some duplicate-' and be missing some observations from *D* with the result that th< values of *mfo* $(f_j \ D_i)$ for each feature 7 j willhkelv b< different on each bootstrap suhsarnple *D,* Under some additional assumptions we then proceed as if the bootstrap samples were obtained from the actual population *P*

Significantly Better uses two different bootstrap methods described by Noreen [1989] the Normal Ap proximation Method, and the Shift Method Figure 3 shows the compulation of some quantities used in the above methods *Diffo t*he difference in information value between $f_{ean}d$ [a]nd *fpref* computed on the set of

Given

N_B (Number of bootstrap samples to generate)

$SigLevel$ (significance level)

Function SignificantlyBetter(f_{cand}, f_{pref}, D)

Generate bootstrap samples $P_B = \{D_1, D_2, \ldots, D_{N_B}\}$ by repeated subsampling with replacement from D

Calculate

$$Diff_D = IDiff(f_{cand}, f_{pref}, D)$$

$$\mu_{P_B} = \frac{\sum_{i=1}^{N_B} IDiff(f_{cand}, f_{pref}, D_i)}{N_B}$$

$$\sigma_{P_B} = \sqrt{\frac{\sum_{i=1}^{N_B}(IDiff(f_{cand}, f_{pref}, D_i) - \mu_{P_B})^2}{(N_B - 1)}}$$

$$PV_N = Prob_{N(0, \sigma_{P_B})}(Diff_P \geq Diff_D) \simeq Prob_Z(Z \geq \frac{Diff_D}{\sigma_{P_B}})$$

$$C_S = Diff_D + \mu_{P_B}$$

$$NGE = \sum_{i=1}^{N_B} GE(IDiff(f_{cand}, f_{pref}, D_i), C_S)$$

$$PV_S = Prob(Diff_{P_B} \geq C_S) \simeq \frac{NGE + 1}{N_B + 1}$$

Where

GE(a,b) is 1 if $a \geq b$ and 0 otherwise

$IDiff(f_i, f_j, D_x) = Info(f_i, D_x) - Info(f_j, D_x)$

If $PV_N \leq SigLevel$ and $PV_S \leq SigLevel$

Then **Return**(true)

Else **Return**(false)

EndIf

Figure 3 Function **SignificantlyBetter**

training data D up$_B$, the mean of our statistic of interest (difference in the information value between *fcand* and *fpref*) over the the bootstrap samples and ap$_B$ the standard deviation of our statistic of interest over the bootstrap samples

The Normal Approximation Method operates under the assumption that the sampling distribution of the statistic of interest (difference in the information value of *fcand and* *fpref)* under the null hypothesis (no difference) is normally distributed with mean zero and same variance as in the bootstrap samples This assumption is used to calculate P V N , the probability under the .Norma/assumption that a value of our statistic higher than or equal to $Diff_D$ could have been obtained by chance To calculate PVn, we use σ_{P_B} and the probability function/tables for a standard normal distribution

The Shift Method assumes that the sampling distribution of the statistic of interest on the complete population *P* has the same shape, but different mean, than the sampling distribution over the bootstrap samples *Pg* To determine the corresponding *PVs,* the probability under the *Shift* assumption that a value equal or higher than *Diffo* could have been obtained by chance we count the number of times that the value of the statistic on bootstrap samples is higher than a shift criterion ($C_S = Diff_D + \mu_{P_B}$), and divide that count by the number of subsamples *NB*

We only decide that the feature f_{eand} is significantly} better than f_{pref} if it is significantly better according to both the Normal Approximation Method and the Shift Method SignificantlyBetter is computationally quite expensive However during the selection of most features this needs to be done verv few times If the feature with the highest initial information value is the feature of highest preference SignificantlyBetter never needs to be computed When other features are Initally selected onh the features with higher preference are checked As soon as one significant difference is computed, no other significance computation is necessary For the sake of efficiency the case where two or more insignificant differences could account for a single significant difference is not considered Our interest is not in precise computations of significance, but rather the qualitative effect of significance testing on the selection of attributes m *CA* 5 while retaining a reasonable level of efficiency

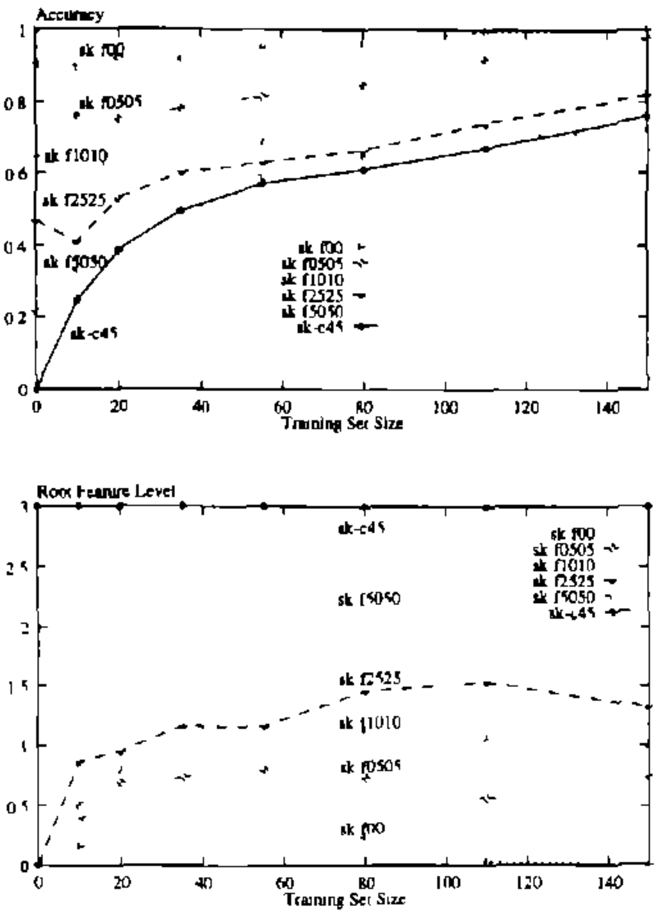
4 Experimental Design

To test our approach we conducted experiments with the audiology dataset from the UCI (University of California at Irvine) Machine Learning repository This database contains 226 examples from 24 classes Each example is described bv 67 discrete features

For each of the 30 learning trials of our experiments a test set of 76 examples and a disjoint training set of 150 examples were randomly and uniformly selected Because we want to Lest the robustness of various strategies in the face of unrepresentative data we sorted the training examples according to their Euclidian distance fiom a randomly selected datum The training set was further divided into subsets which contain the first 10, 20 35, 55 80, 110, and 1 50 (I e , all) examples of the sorted training set Decision trees are learned for each of these subsets of training data

The dependent variables of interest are predictive accuracy and the level (as illustrated in Figure 1) of the root feature of the decision trees learned under different conditions Due to the recursive nature of decision tree induction we expect that the tendencies observed at the root can be extrapolated to other nodes of the tree

The independent variables are the size of the training set, the significance level used for hypothesis testing in our variation of (*A* 5 and the degree of model imperfection Note that while varying the size of the training set we are also varying its degree of skewedness because training data are ordered based on Euclidean distance, small samples lend to be drawn from a small portion of the data description space, the larger the training dataset, the higher the proportion of the complete data set it covers, and thus the more representative it becomes For the largest data set of 150 examples all skewedness disappears since all 150 examples were ran-



domlv chosen from the complete data set W< present results from skewed sampling as this tends to repp sent worst case conditions for our learning system We have also experimented with traditional sampling (for all training sizes) thus allowing us to better tease apart the influence of skew and training set SIZE though this paper does not elaborate on this issue Significance levels of 50% 25% , 10% and l% are varied to indicate increasing confidence in the quality of our models

We follow Mooney s approach [Moonev 1993] for generating theories of varying degrees of imperfection A perfect theory, 1 e , one that correctlv, classifies 100% of all audiology examples was first constructed b> running C 4 5 on the complete data set of audiology examples with all pruning disabled This theory (named *fOO)* contained 86 rules with an average of 7 79 antecedents per rule Imperfect theories (named *f55 flOlO f2525 and f5050)* were generated bv randomlv adding and then randomly deleting a percent of all conditions from the rules of the perfect theory (a corresponding *5% 10% 25% and 50%)* This results in contaminated theories with errors both of omission and of commission The accuracies over the complete data set of the imperfect theories *f55 f1010, f2525* and/5050 are *89%, 65% 46%* and *21%,* respectively For comparison, the accuracy obtained if we always predict the most frequent class in the dataset is *21%*

5 Experimental Results

Figure 4 2 shows results from a baseline study The curve labeled *sk-c45* shows the results of standard C4 5 on skewed trials with 'raw' features only The rest of the curves in this figure show the accuracv and root feature level for C4 5 when model features are added to the description of the data, but no preference ordering or hypothesis testing is done We can see that accuracv improves significantly (over *sk-r45)* when a domain theory is exploited even for a low quality theory such as f*5050* These results compare favorably, to other systems tested on this domain [Mooney 1993}] [Ourston 1991]

An interesting fact illustrated in Figure 4 is that the number of training examples required Tor C4 5 with a domain theory *{sk-fxx* curves) to produce an accuracy equivalent to the corresponding theory alone seems to be inversely proportional to the qualitv of the theory the two extremes being *f5050* (0 training examples required to reach 21% accuracy) and *fOO* (all training examples required to reach 100% accuracy) Ideally however (he accuracy of our system should be equal or better than the accuracv of the model alone, or the C4 5) learning algorithm alone This only occurs for large enough or representative enough training data sets This behavior is not unique to our system it can be observed in the published learning curves shown for some systems that combine analytical and empirical learning[Pazzani and Kibler 1992] [Ourston 199l] As we will see significance testing of ranked features appears to mitigate this undesirable behavior

Figure 4 gives a good indication of the tvpe of features, selected with theories of varying quality. Standard C4 5 can only access raw features (level 3) C4 5 with features from a perfect modf 1 (1 e *fOO)* chooses almost exclusively (hut not always) the model prediction feature (level 0) With lower quality theories C 45 gradually chooses features of greater level numbers However for the perfect model, there should be no reason to choose anv feature other than the model prediction feature This does not happen due in part to a known bias of the information gain ratio against features with many values (In the audiology domain the model prediction feature has 24 possible values, other model features are binary and raw features have few values) However as We will see next this problematic bias can be mitigated with the use of significance testing

Figure 5 shows the effects of significance testing of ranked features when C45 is augmented with features from a perfect model Rather than the 150 examples

2Tables corresponding to (he graphs in this paper ron laming means and standard deviations ean 1M found at http //www *vuse* vanderbilt edu/^dfisher/iech reports/ijcai95-tables pa

3 Here we use standard 045 for learning For testing wt use the classlfiration procedure described in section 2 I e predicting with the model rather than the most frequenl class at leaves of the decision tree where there is insuffieirnt data for further decomposition

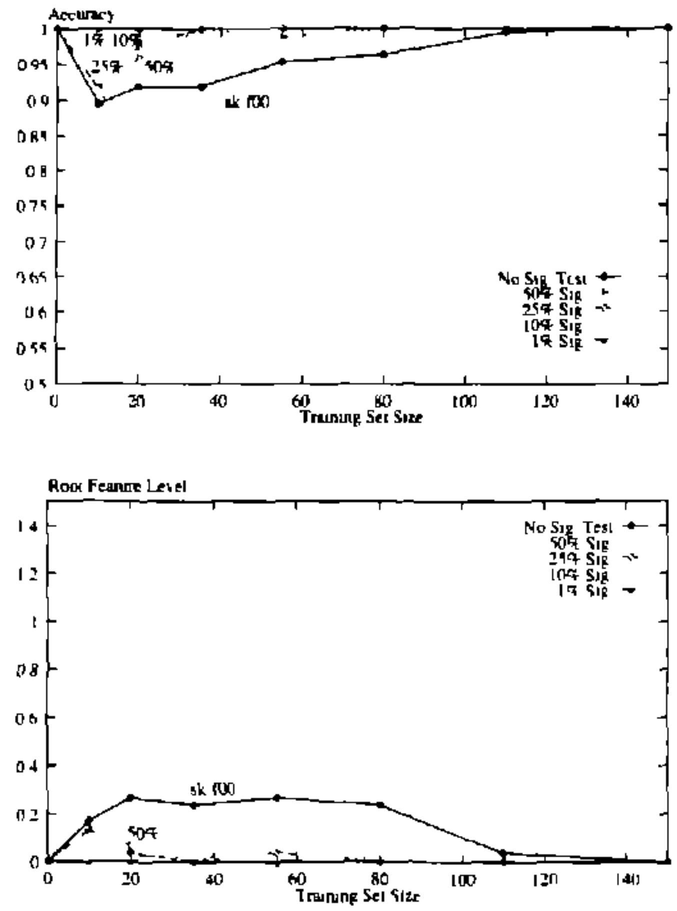


Figure 5 Average acruracv and root feature level of decision trees learned with features from a perfect model under varying levels of significance

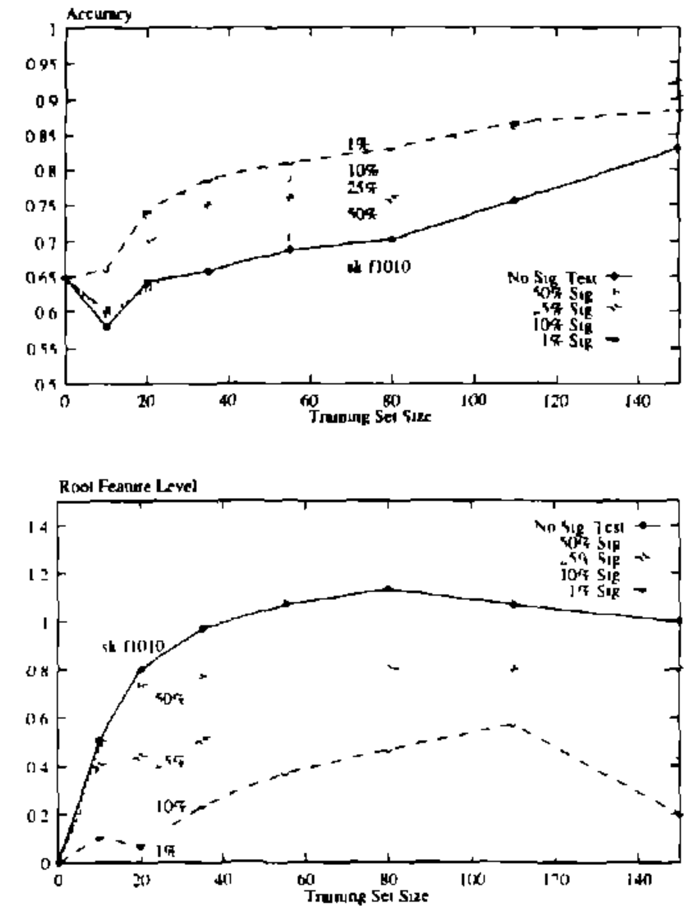


Figure b Average accuracv and roof feature level of deci sion trees learned with features from the *flOlO* imperfect model under varving levels of sigr

needed before hypothesis testing of ranked features results in 1009c accurarv with only 80 examples when using a 509c significance level or with just 85 examples when using more strict significance levels (259% 10% or 1%) This figure also shows how the average root featur< level is gradually} reduced with stricter significance levels

Perfect domain theories are an interesting boundary, case, but most interesting theories are imperfect The graphs of Figure b illustrate how stricter levels of sig nificance achieve our objective of biasing C45 toward features of smaller level number using the *flOlO* domain theory This figure also shows that the accuracy obtained with the *flOlO* imperfect theorv improves consistently with stricter levels of significance testing (509c *40%* 10%) for any size of the training set, including the complete training set of 150 examples In addition, while with no significance testing (or the almost equivalent significance testing at the 50% level) at least 55 examples are needed to obtain better accuracies than the *flOlO* theorv alone with stricter levels of significance testing only JO examples are required For the 1% significance level, accuracy, is better than other significance levels when the training sets contains less than 110 examples, and worse when the training set contains more than 110 examples Thus at least in this domain there is a breakeven point for significance levels that depends both on the quality of th< model and the size of the training set after which stricter significance levels are detelmental In our experiments the size of the training sets corresponds to increasing quality in t he available data in the sense that thev are better representatives of the complete population both due to the sheer amount of data and due to the fact that the skewedness effect we introduced in the training set tends to dimmish as tht size of the trainings sets are increased

For lower quahtv theories similar behavior is observed with increasingly strict significance levels, accuracy improves when the training stt contains few examples, and decreases with training sets that contain manv examples Thus, for every combination of model quality (l e amount of contamination in the model) and data quahtv (level of skewedness and number of training examples) there is an optimal level of significance between the extremes of 50% and 0% in this domain However the choice of a beneficial but perhaps non-optimal sig nificance level is not difficult Significance levels only seem to become detrimental for large data sets when we use significance levels stricter than 10% Values between

50% and 10% always seem to improve accuracy (perhaps non-optimally) for any size of the training set Thus, expert intuition about the trustworthiness of an existing model with respect to the available can be incorporated into our learning algorithm to obtain additional perfor mance improvements

6 Concluding Remarks

In this paper we address a situation that we believe to lie of practical interest learning whenever we have a model believed to be of good quality but imperfect nevertheless together with a set of data of unknown rep rescntativeness and quality We present a method that attempts to take advantage of both the model and the data plus our prior knowledge about the quaht\ of the model Our mc thod biases empirical learning in a flexible manner such that model-based features, or more generally preferred features bised on some a prion determined preference ordering are selected unless sufficient refuting evidence appears in the data The amount of evidence required is determined bv statistical significance and is set by the user according to his/her confidence in the quality of the available model

Our expenniental results show that when features gtn erated from a model are smply added d to the description of the data aecuracy is increased to a degree proportional to the quality of the model However sum*" problems with this simple approach are illustrated by the fact that perfect models only result in perfect accuracy with large or very representivetve srts of traning exam ples if significance testing with a preference, ordering is used with a pe rfect model our system heroines more robust in the presence of skewed data few examples are then needed 1o obtain perfed accuracy Further with imperfect models of good quality we obtain additional increases in accuracy for any number of training exam pies

Although significance testing lias been used previously m machine learning methods such as for the pruning of decision trees [Qumlan]9t<b] our use. of Ibis concept for flexibly introducing prior knowledge bias in empirical learning seems to be novel

Acknowledgements

This research was supported bv a grant from NASA Ames Research (enter (NAG 2-834) to Doug Fisher We thank Deepak Kulkarrm and Peter Robinson for early dis cussion on initial strategies that led to those described here

References

[Clark and Malwm 1993] P (lark and S Matw in Using qualitative models to guide inductive learning In Proceedings of the Tenth International Conference on Machine Learning pages 49-5b Amherst, MA 1993

[Draslal et al , 1989] G Drastal, G Czako and S Raatz Induction in an abstraction space a form of constructive induction In Proceedings of the Eleventh International Joint Conference on Artificial intelligence, pages 708-712 Detroit MI, 1989

[Efron and Gong, 1983] Bradley Efron and Gail Gong A leisurely look at the bootstrap the jackkmfe and cross-validation The American statistician 37(1) 36-48,1983

[Evans and Fisher, 1994] R R Evans and D Fisher Overcoming process delavs with decision trer induction IEEE Expert, 9(1) 60-66 1994

[Mooney 1993] R J Mooney Induction over the une\ plained Using overly-general theories to and concept learning Machine Learning 10 79—110 1993

[Musick et al, 1993] Ron Musick Jason (allell and Stuart Russell Decision theoretic suhsampling for induction on large databases In Proceedings of the Tenth International (onference on Machine Learning pages 212-219 Amherst, MA 1993

[Noreen 1989] Eric W N oreen Computer Intensive Methods for Testing Hypotheses An Introduction John Wiley K. Sons New York NY 1989

[Ortega 1994] J Ortega Making the most of what you ve got using models and data to improye learning rate and prediction accuracy Technical Report (S-94-01 Computer Selence Dept \anderbilt University Abstract appears in Proceedings of the Twlfth National (onferece nce on Artificial Intelligence p 1483 Seattle \\A 1994

[Ortega In Preparation] J Ortega Making the Most of What YOU Got using Models and Data to Improve Prediction tecuracy PhD thesis. \ anderbilt Univer sity Nashille TN In Preparation

[Ourston 1991] D Ourston I stng Erplanation-Bastd and Empincal Methods in Theory Reaision PhD thesis Iniversitv of Texas Austin, 1 \ 1991

[Pazzatn and Kibler 1992] M Pazzani and D Kibler The utility of knowledge in inductive learning Ma chine Learning 9(1)37-94 1992

[Qninlan 198b] 1 R Quinlan Induction of decision trees Machine learning 1 81 10b 198f>

[Quinlan 199i] I R Quinlan < 4 5 Programs for Ma (hine Learning Morgan Kaufmaim San Maleo C \ 1993

[Rendell and Seshu 1990] Larry Rendell and Raj Seshm learning hard concepis through constructive induction framework and rationale Computational Intelligence 6(5) 247 270 1990

[Robinson 1993] Peter Robinson Auitomated fault diagnosis algorithms for the reaction control system of the space shuttle Technical Report FIA-93-05 N A S A . Ames AI Research Center 1993

[Towell et al 1990] G G Towell J \\ Shavlik and M O NSoordewier Refinement of approxmate domain theories by knowledge-hased in ural networks In Proceedings of the Eighth National (onfertice on \rtificial Intelligenct pages86l-8Cb Boston, MA 1990