Are We Losing Accuracy While Gaining Confidence In Induced Rules - An Assessment Of PrIL

F. Özden Gür Ali

GE Corporate Reseach and Development
K1-4C29A, P.O. Box 8, Schenectady NY 12301
e-mail: gurali@crd.ge.com

William A. Wallace

Decision Sciences and Engineering Systems Department
Rensselaer Polythecnic Institute
15th Street, Troy, NY 12180
e-mail: wallaw@rpi.edu

Abstract

Probabilistic Inductive Learning (PrIL), is a tree induction algorithm that provides a minimum correct classification level with a specified confidence for each rule in the decision tree. This feature is particularly useful in uncertain environments where decisions are based on the induced rules.

This paper provides a concise description of (the extended) PrIL and demonstrates that its performance is as good as best results in the machine learning literature, using datasets from the data repository at UC Irvine.

Introduction

One objective of database mining is to discover knowledge that will be used to solve problems or make decisions; i.e., it is prescriptive in nature. In that case it is our contention that we should provide the user with measures of goodness as well as the uncertainty of the prescriptions. For example, in database marketing we would like to be able to have a 90% confidence that if we mailed out 100 promotions at least 80 of them reach the target population.

Our approach, Probabilistic Inductive Learning (PrIL), uses a tree induction algorithm and provides a minimum correct classification level with a specified confidence for each rule in the decision tree. The purpose of this paper is to provide a concise description of PrIL and compare its performance to other machine learning algorithms using datasets from the data repository at UC Irvine.

Probabilistic Inductive Learning

PrIL consists of two parts: 1) an initial branching structure where a complete tree is spun using chosen attributes, and 2) an independent subset elimination component for each branch where rules are posted. PrIL branches enough to account for the important attribute dependencies and main effects without diminishing the sample size to such an extent that probabilistic statements cannot be made. A previous version of PrIL has been published (Gür-Ali & Wallace 1993). This version adds levels of reliability for rules, replaces the chi-square approximation for very high reliability levels by Poisson approximation and delineates the parameters of the algorithm.

Figure 1 shows a hypothetical PrIL tree where A_i are the attributes that describe a case; the boxes with different shadings represent the classification categories. The box with the question mark indicates that no rule applies to such cases. The numbers on the arcs correspond to the values of the attributes in the parent node and indicate the branch to follow according to the attribute value of a case. The numbers in italics next to the boxes are the goodness measures of the rules: the correct classification proportion for the rule will equal or exceed this figure with a stated confidence level.

There are two types of nodes in the PrIL tree. The total branching nodes, represented by small circles, have only non-leaf nodes as descendants. Their number of generations is fixed, they can be parents of subset elimination nodes which are represented by ovals in the figure. A subset elimination node cannot have a total branching node as descendant. Both types of nodes have only one parent. Subset elimination nodes are characterized by the fact that they have up to one non-leaf descendant.

Taking the right-most branch in Figure 1. would establish the following rule: "If attribute 1 has value 2, and attribute 2 has value 2, and attribute 6 does not have value 1, and attribute 4 has value 4 then classify the case into the category with dots."

The aim of PrIL is to provide decision rules that individually satisfy the minimum reliability requirements set by the user with a prespecified confidence. Categories may have different amounts of risk associated with them giving rise to different maximum tolerable misclassification levels. For example, the user may want to treat "giving a loan to a marginal applicant" differently from "not giving a loan to a qualified applicant".

Categorization

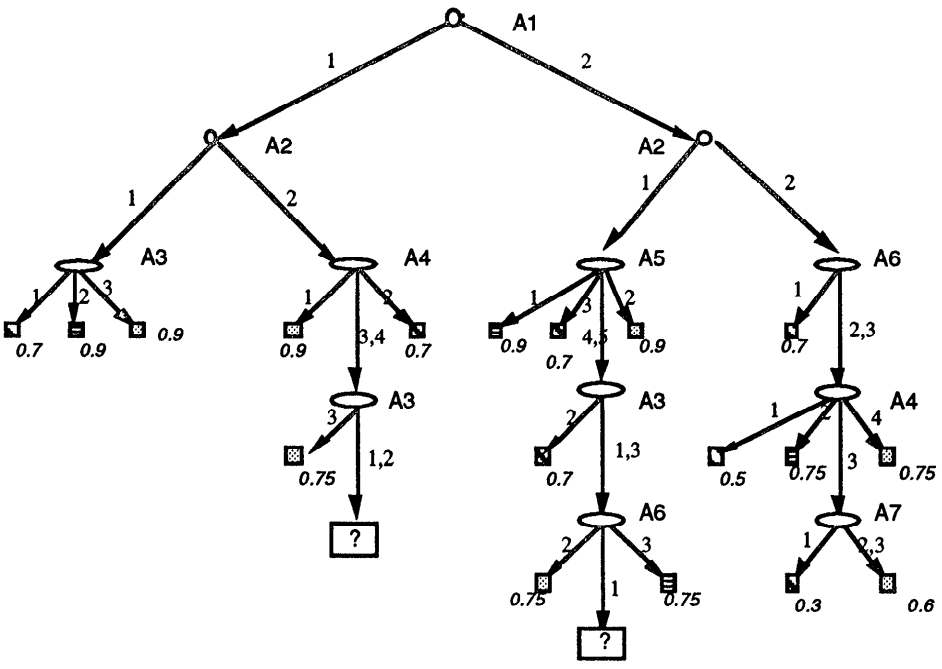
PrIL uses the abundance of data typically available. Continuous variables are categorized at the beginning of induction and the categorization is maintained throughout the process, unlike other methods that recategorize at every iteration (Breiman et al. 1984) (Quinlan 1987) (Carter & Catlett 1987). The result is an unchanging definition of ranges on the categories which facilitates communication of the rules to potential users. The obvious disadvantage is a possible loss of accuracy. The variables are categorized based on their univariate distribution only. Although other algorithms like ChiMerge (Kerber 1992), D-2 (Catlett 1991) or (Fayyad & Irani 1992) take the classification variable into account in categorization, there is no guarantee good results will be achieved when interactions with other attributes are present. Assuming no interaction effects is quite naive in most practical problems.

Modeling / Selection Of Branching Attributes

As can be seen from Figure 1, the PrIL tree can be considered as consisting of a number of independent subtrees, one for each value combination of the branching attributes. The branching attributes in Figure 1 are A1 and A2. Their purpose is to utilize the background knowledge about the problem. Modeling refers to identifying attributes that are candidates for the branching phase. The effect of the attribute(s) in the branching phase is taken into account in every rule. Therefore they should be significant main effects or be one of the partners in most of the interaction effects. Attributes that are known to be important become candidates for branching attributes. Another avenue of generating candidates is by fitting logit models to the data and identifying the most significant main effects. Detailed discussion of logit models can be found in (Christensen 1990).

Minimum Reliability Requirements

The tree resulting from using PrIL has a prescriptive value, i.e. rules will be used to guide or make a decision. Therefore, the costs (benefits) of an incorrect (correct) decision must be considered. This information is implemented in terms of minimum reliability levels, R_k. Setting R_1=0.90 would mean that we cannot tolerate any rule that classifies a case into category 1 to be wrong more than 10 percent of the time. When no rule can be developed that meets the minimum reliability requirements for a subset of cases, the cases are called "undecided". There is value in knowing which parts of the problem space we are able to make reliable enough classifications, and where there is insufficient evidence. These requirements can be different for different categories. The misclassification costs and correct classification benefits can be used to calculate them.



LEAVES WITH CATEGORY LABELS
TOTAL BRANCHING NODES
SUBSET ELIMINATION NODES

Figure 1. Tree structure of the Probabilistic Inductive Learning approach

Subset Elimination

Subset elimination tries to find subsets of the data where one category has more than the required proportion of examples with the stated confidence. If multiple subsets are found the one that gives the highest confidence becomes the rule to be posted and all examples covered by it get eliminated from further consideration. The required reliability levels, r_k's, are set at high levels and successively reduced to the minimum required reliability levels, R_k's. The subsets are defined by the values of an attribute or combination of attributes. The maximum number of attributes in a rule, M, is a parameter indicating the specificity versus simplicity of the tree. Simpler rules are given precedence over ones with more attributes. In practice we have not gone beyond M=2.

The so-called "reliability function" dictates the reductions in the reliability requirements of categories at each step. There is no restriction on the shape of this function imposed by the PrIL method, as long as it is monotonic and includes the points (R_1, R_2,..., R_K) and (1, 1, 1,..., 1). For a discussion about the effect of different shapes and number of points on the function, as well as equality of Rk's refer to (Gür Ali 1994).

For each set of reliability requirements $\{r_k\}$ we employ the following process : first a rule with one attribute is searched for, if there is no attribute that satisfies the current requirements, rules consisting of combination of two attributes are considered. This process is continued until all combinations of M attributes are considered.

The next section describes the subset elimination for given $\{r_k\}$ and one attribute in each rule, i.e., M=1. Without loss of generality, the combination of many attributes can be regarded as one attribute with the value consisting of the combination of the individual values of the original attributes.

Subset Elimination For A Given $\{r_k\}$

The subset elimination is applied to each branch independently. We call all cases in the training set that fall to the branch, the subset.

While there are cases in the subset for each attribute A_i, *build the rule set of* A_i and *compute its significance* $S(A_i)$; post the rule set with the maximum $S(A_i)$ and eliminate all cases that are covered by the posted rule set.

The process is repeated until no case is left or no attribute gives rise to a rule.

Letting A_i be the attribute i; v_{ij} be value j of attribute i; Y be the classification variable; c_k be classification category k, f_{ijk} be frequency of cases with $A_i=v_{ij}$ and $Y=c_k$; we *build rule set of* A_i *(*Y_i*)* as follows.

Construct the frequency table that shows the case counts for each value of attribute A_i versus classification categories. For each value v_{ij} of attribute A_i and each c_k do the following

If $f_{ij.} > n$, where $\Sigma_k f_{ijk} = f_{ij.}$, then :

Test $H_0 : P_{ijk} \geq r_k$ with a significance level of α, where

P_{ijk} : population parameter for $P(Y = c_k, A_i = v_{ij})$, and

r_k : current required reliability level for classifying cases to c_k.

If H_0 cannot be rejected at a level then add the rule "if $A_i = v_{ij}$ then classify the case to $c_{ij}*$" to the rule set of A_i, where $c_{ij}*$ denotes the category that the rule classifies the cases into. In order to prevent the building of rules based on very small sample sizes if fij.<=n no rule ρ_{ij} is established. f_{ijk} is multinomially distributed given $f_{ij.}$. Since we are only interested in the number of cases in category c_k versus the rest, we can model the distribution of f_{ijk} given $f_{ij.}$ with binomial distribution

$$F(f_{ijk}) = \sum_{m=0}^{f_{ijk}} \binom{f_{ij.}}{m} P_{ijk}^{m} (1- P_{ijk})^{f_{ijk} - m}.$$

For H_0 the uniformly most powerful test is of the following form: Reject H_0 if $f_{ijk} < \text{const}$;

with $P(f_{ijk} < \text{const} \mid P_{ijk} = r_k) = 1 - \alpha$.

To *compute* $S(A_i)$ we let $f_{ij}*$ denote the case count with $A_i = v_{ij}$ and $Y=c_{ij}*$, and $r_{ij}*$ be the required reliability associated with the rule of v_{ij}, as shown in the table. Y_i denotes the rule set of the attribute A_i. The rule associated with v_{ij} is called r_{ij}. Since $f_{ij}*$ is binomially distributed with $P_{ij}*$ and $f_{ij.}$, as fij. increases the binomial approaches the normal distribution with mean $f_{ij.}P_{ij}*$ and variance $P_{ij}*(1-P_{ij}*)f_{ij.}$ (Rohatgi 1976).

ρ_{ij} in Ψ_i	$c_{ij}*$	not $c_{ij}*$	required reliability
v_{i1}	$f_{i1}*$	$f_{i1.}-f_{i1}*$	$r_{i1}*$
:			
v_{ij}	$f_{ij}*$	$f_{ij.}-f_{ij}*$	$r_{ij}*$
:			

Table 1. Frequency table for rules in Ψ_i from the lack of fit point of view.

Therefore, $\dfrac{(f_{ij}* - f_{ij.}\, P_{ij}*)^2}{f_{ij.}\, P_{ij}* (1-P_{ij}*)}$ is asymptotically χ^2_1 distributed

and $\sum_{\rho_{ij} \in \Psi_i} \dfrac{(f_{ij}* - f_{ij.}\, P_{ij}*)^2}{f_{ij.}\, P_{ij}* (1-P_{ij}*)}$ is asymptotically $\chi^2_{\Psi_i}$ distributed.

Under the null hypothesis that $P_{ij}* = r_{ij}*$,

$X^2 = \sum_{\rho_{ij} \in \Psi_i} \dfrac{(r_{ij}* f_{ij.} - f_{ij}*)^2}{r_{ij}* f_{ij.}(1-r_{ij}*)}$ would follow the $\chi^2_{\Psi_i}$.

All rules of the rule set of A_i must have passed the test $P_{ij}* > r_{ij}*$, therefore, all $f_{ij}* > r_{ij}* f_{ij.}$ and X^2 shows the divergence from the required reliability in the direction of less misclassification. Therefore, its significance $S(A_i) = F_{X^2}(X^2)$ where $F_{X^2}(.)$ denotes the cdf of chi-square

distribution with the corresponding degrees of freedom. When $r_{ij}* >= 0.99$, we use the Poisson approximation to binomial to find the significance of the rule set.

The PrIL algorithm may leave a subset of the training cases not covered by any rule. They are labeled as undecided and are essential to understanding the extent to which we can confidently predict classification.

The goodness of a PrIL tree can be judged without resorting to test sets. An overall reliability measure for category k can be obtained by taking the weighted average of the reliability levels of rules that classify the cases they cover into category k. If we are interested in the overall reliability as opposed to category specific reliability the average reliability will be the appropriate measure. One way of taking the undecided percentage into account is to use the adjusted average reliability, where an undecided case is accounted as if it were misclassified.

Comparison Of PrIL Performance To Other Machine Learning Algorithms

The distinguishing feature of PrIL is that it provides goodness measures for individual rules. To show that this feature is not attained at the cost of accuracy we have chosen six datasets from the data repository in UC Irvine (Murphy & Aha 1992). These datasets have served as test data for a variety of algorithms, hence they facilitate a comparison between those algorithms and PrIL. They represent a wide spectrum of problem domains and data characteristics. Like many other classification algorithms, PrIL has a set of parameters that need to be set before trees can be induced. For example, to use CART we need to choose the rule for best split: twoing criterion or GINI index; for C4 we need to select the pruning procedure. However, when setting PrIL parameters the user incorporates his/her values such as risk toleration on each category, background information on important attributes and the desire of more complicated versus simpler trees.

Changing the parameters results in changes in the percent of undecideds, the simplicity of the tree, the weight given to the accuracy of one category versus others and the confidence we have in the correct classification estimates for rules.

In the machine learning community, accuracy of classification appears to be the criterion that is used to compare algorithms. Since the norm is to classify all the cases, there is no consideration for undecided cases. We will assume that the undecided cases that are produced by PrIL are misclassified and will define accuracy as (number of correctly classified cases/number of all cases in the test set)*100, and call it adjusted accuracy. Note that this is a very conservative estimate of accuracy.

The three approaches used to select the PrIL parameters are as follows.

a) Choose the settings that lead to the highest mean *accuracy* levels from five *test sets*. This appears to be the common practice in the literature although it is optimistically biased because it uses the same test set to select parameters and evaluate the accuracy of the tree.

b) Choose the settings that produce the highest mean of *adjusted average reliability* levels from five *training sets*. Adjusted average reliability is obtained from the training set and is the weighted average of the reliability levels of the rules, where the number of cases constitutes the weight and the undecideds are treated like a rule with zero reliability. The accuracy of the trees induced with the chosen parameter settings are evaluated on the test sets.

c)Choose the setting that are recommended for high adjusted accuracy.

The experiment is designed in the following way: For each data set 5 training and 5 test sets are generated;l each combination of PrIL parameters are run on all training sets and evaluated on the corresponding test sets. The data sets are the following. 1) LED display, 2)LED+17, 3)Waveform, 4)Mushroom, 5)Congressional Voting Records, and 6) Credit Approval (Murphy & Aha 1992). The continuous values in credit application and waveform datasets have been categorized before applying the PrIL algorithm, based on the quartiles of the empirical distribution.

For datasets 3-6 we generated the nth dataset by putting every ith case for which i mod 5 is n, into the test set and keeping the rest in the training set. For the LED display and waveform data domains 5 training sets of 1000 cases and 5 test sets of 500 each were generated using the programs from the data repository.

In machine learning literature, accuracy is measured in essentially four ways :1) *Resubstitution estimate*, which is optimistically biased; 2) *Test set estimate*, with varying test sample sizes stratification practices; 3) *V-fold validation*, which provides both a location and variability estimate for accuracy; and 4) *Generation of multiple training and validation data sets from the known model*, which requires the knowledge of the true model.

In our evaluation, we used v-fold validation with v=5 replications for the mushroom, vote and credit data sets, and generated 5 new training and 5 new test data sets for the waveform, LED and LED+17 data domains. We report the mean adjusted accuracy (as defined above) and its standard deviation. Adjusted accuracy is a conservative measure in the sense that it counts undecideds as misclassified.

The parameters and their possible values, enclosed in brackets, are as follows.

1) M, maximum number of attributes in a subset elimination rule {1,2},

2) n, minimum number of cases required to post a rule {5, 10. 20},

3) shape of reliability function {convex, linear, concave},

4) symmetry of reliability function in classification categories {yes, no};

5) number of steps in the reliability function {3, 5};

6) number of attributes in the branching module {0, 1, 2}, and

7) n option, i.e., option to disregard n for undecideds {0, 1}.
The recommended parameter settings for high adjusted accuracy are : M=2; n=5; linear and symmetric reliability function with 5 steps; and disregard n for undecideds. The number of branching attributes was determined by logit analysis as the number main effects significant at 0.05 level : one branching attribute for mushroom, vote and credit datasets, none for waveform, and two for LED and LED+17. In general, the most significant main effects in the logit analysis were chosen as the branching attributes (Gür Ali 1994).

Results

The results of the experiment are shown in Table 2. The adjusted accuracy of the test sets is the measure to be compared with the results from literature. For mushroom data many trees achieved 100% adjusted accuracy.

For congressional vote data according to the adjusted average reliability we choose the following parameter settings, M=2 , n=5 with 5 steps, no branching attribute, and n-option on. These parameter values produce trees with an average of 13.6 rules. The mean adjusted accuracy turns out to be 94.02% with standard deviation of 0.96. Note that these settings are not the best in terms of the adjusted accuracy. The five trees that turned out to be most accurate have very few rules (an average of 2.8 rules). There are five different sets of settings of PrIL that produce the same highest adjusted accuracy for the vote dataset. They all have M=1, n=20, one branching attribute and n option off, producing broad rules.

The best settings for the credit dataset produce a mean of 17.4 rules, the standard deviation of the adjusted accuracy is very low, 0.89, while the mean adjusted accuracy of 86.98 is the best result in the literature.

The best trees from the waveform data have considerable number of rules, and have M=2. This reflects the complex nature of the waveform data. The best adjusted accuracy on waveform data is 76.48 with a standard deviation of 2.44.

It is interesting to note that the LED and LED+17 datasets have about the same adjusted accuracy, showing that PrIL is able to handle noise. But in general, the trees induced from LED+17 data have more rules than those from the LED data. We also notice that although the standard deviation of adjusted accuracy for LED+17 data is not more than LED data, the standard deviation of the number of rules is higher. Hence, noisy data shows itself in the tree structure rather than the accuracy.

We compared adjusted accuracy, column 1, with test results from literature, column 4. We have conducted t-tests with an a level of 0.05 to see if there is a significant difference between these adjusted accuracy figures and the best accuracy figures in literature. It turned out that except for LED dataset they were not significantly different.

We also compared PrIL results, column 3, obtained using recommended settings, with best results from the literature, column 4. We used a two tailed t-test with a = 0.1. It turns out that all PrIL accuracies were at least as good as the best in literature, and that for credit dataset PrIL performance was significantly better.
In general we see that PrIL algorithm has performed very well compared with other algorithms. Table 2 includes the range of accuracies reported in the literature for each dataset.

We have seen that the recommended settings for the PrIL parameters give better accuracy figures than the settings chosen by evaluating all possible parameter settings according to their average adjusted reliability on the training set, which is computationally expensive. Therefore, we can simply use the recommended settings for best results in adjusted accuracy.

	1		2		3		4
Selection Criterion	Adjusted Accuracy		Adjusted Average Reliability		Recommended Settings		Reported Accuracy Range in Literature
Data	Mean	Std	Mean	Std	Mean	Std	
mushroom	100	0	100	0	100	0	95 - 100
vote	96.22	2.72	94.02	0.96	94.02	2.12	90 - 95.6
credit	86.96	0.89	82.32	3.72	82.61	1.98	84.4
waveform	76.48	2.44	70.68	2.23	76.48	2.44	72 - 78
LED	70.80	2.72	69.20	1.50	69.20	1.50	71 - 74
LED+17	70.64	1.75	70.48	1.78	70.64	1.75	41 - 71.5

Table 2. Mean and standard deviation of adjusted accuracy on the test sets obtained by PrIL, and the range of accuracies reported by other algorithms. The literature referred to are (Breiman et al. 1984), (Murphy & Aha 1992), (Schlimmer 1987), (Holte 1993) and (Anderson & Matessa 1992).

Conclusions

PrIL classification algorithm has been designed to address real life problems with following characteristics :

- Large number of examples, expressed as n-tuples, representative of the variety of cases and their frequency of occurrence is available.
- There is considerable uncertainty involved in the system, such that two examples with the same attribute-values can have two different classifications.
- The misclassification costs or correct classification benefits are not necessarily the same for different classification categories.
- It is desirable to assess the uncertainty of a rule in terms of the percent of time we can expect the rule to classify correctly for a given confidence level.
- The decision process is to be automated only as far as warranted. Trying to classify cases with low reliability would reduce the quality of the decision process.
- It is desirable to gain insight into the system and explain why a certain decision is recommended or a certain outcome is predicted.

In this paper we have demonstrated that PrIL provides us with accuracy comparable to the best accuracy in literature. For all datasets, the best performance of PrIL is either superior to or not significantly different from best results reported in the research community. Our objective is not to declare that PrIL is superior in accuracy but to show that we are not losing in accuracy while we are gaining reliability measures for individual rules.

References

Gür-Ali, Ö., and Wallace, W. A. 1993. Induction of Rules Subject to Quality Constraints: Probabilistic Inductive Learning. *IEEE Trans. Knowledge and Data Eng.* 6(20):979-984.

Breiman, L.; Friedman, J.H.; Olshen, R.A., and Stone, C.J., 1984. *Classification and Regression Trees.* Belmont, CA : Wadsworth Int.

Quinlan, J. R., 1987. Decision Trees As Probabilistic Classifiers. In Proc. 4th Workshop on Machine Learning; Langley, P., Ed., Los Altos, CA: Morgan Kaufmann.

Carter, C., and Carlett, J. , 1987. Assesing Credit Card Applications Using Machine Learning. *IEEE Expert,* 2(3): 71-79.

Kerber, R., 1992. Chimerge : Discretization of Numeric Attributes. In Proceedings of th 10th National Conf on Artificial Intelligence. AAAI Press

Catlett, J., 1991. On Changing Continuous Attributes into Ordered Discrete Attributes. In Proc. European Working Session on Learning

Fayyad, U. M., and Irani, K. B., 1992. On The Handling of Continuous-Valued Attributes in Decision Tree Generation. *Machine Learning* 8:87-102.

Christensen, R., 1990. *Log-Linear Models,* New York : Springer Verlag,

Rohatgi, M., 1976. *An Introduction to Probability and Mathematical Statistics,* New York : John Wiley.

Gür Ali, Ö., 1994. Probabilistic Inductive Learning : Induction of Rules With Reliability Measures for Decision Support. Ph.D. diss., Decision Sciences and Engineering Systems Department, Rensselaer Polytechnic Institute.

Murphy, P. M., and Aha, D. W., 1992. *UCI Repository of machine learning databases* Machine-readable data repository. Irvine, CA: University of California, Department of Information and Computer Science.

Schlimmer, J. C., 1987. Concept Acquisition Through Representational Adjustment. Ph.D. diss. Department of Information and Computer Science, UC Irvine, CA.

Holte, R.C., 1993. Very Simple Classification Rules Perform Well on Most Commonly Used Data sets. *Machine Learning,* 11:63-91,

Anderson, J.R., and Matessa, M., 1992. An Incremental Bayesian Algorithm for Categorization. *Machine Learning* 9: 275-308.