Learning with Globally Predictive Tests

Michael J. Pazzani

(pazzani@ics.uci.edu)
Department of Information and Computer Science
The University of California
Irvine, CA 92697
949-824-5888

Abstract. We introduce a new bias for rule learning systems. The bias only allows a rule learner to create a rule that predicts class membership if each test of the rule in isolation is predictive of that class. Although the primary motivation for the bias is to improve the understandability of rules, we show that it also improves the accuracy of learned models on a number of problems. We also introduce a related preference bias that allows creating rules that violate this restriction if they are statistically significantly better than alternative rules without such violations.

1 Introduction

A variety of rule learning systems have been developed that create rules to predict class membership of examples such as AQ15 [1], CN2 [2], ITRULE [3], C4.5-rules[4], FOIL [5], FOCL [6], Greedy3 [7], Ripper [8], and decision lists [9]. One commonly reported advantage of modeling predictive relationships with rules is the comprehensibility of the learned knowledge. Rule learners produce a set of learned rules of the form:

$Test_1$ & … & $Test_n \longrightarrow Class_i$

where each test compares an attribute A_i to a value V_{ij} for that attribute. For nominal attributes, the possible tests include determining whether an attribute value of an example is equal to a particular value, is not equal to a particular value, or is a member of a set of values. For numerical values, the tests will determine whether an attribute value of an example is greater than or less than a particular value. Typically, the rules are ordered so that to classify an example, one predicts the class of the first rule whose antecedent is true. One common approach for ordering rules is an estimate of the accuracy of the rule (e.g., Quinlan [4]; Clark & Niblet [2]; Ali & Pazzani [10]).

Table 1 shows an example of some rules learned to screen infants for mild mental retardation [11] from a sample of over 4000 examples collected by the National Collaborative Perinatal Project of the National Institute of Neurological and Communicative Disorders and Stroke. The rules are relatively easy for an expert or novice to understand and could easily be applied by a person or a computer. However, the rules contain certain tests that are counter-intuitive and puzzling to experts. In particular, the third rule predicts that there is low risk of mental

Table 1. Some rules learned to screen infants for mild mental retardation.

```
IF the child has no emotional problems
AND the mother has normal IQ
THEN the risk is LOW

IF fetal distress is ascertained prior to or during labor
AND the mother's education level is less than 12 years
AND the mother smokes
THEN the risk is HIGH

OTHERWISE IF the child has no emotional problems
AND the mother's education level is at least 12 years
AND there were previous stillbirths
THEN the risk is LOW
```

retardation and contains a condition “there were previous stillbirths” that is normally thought of a risk factor for mental retardation. It is possible that this rule is a new medical finding for a sub-population of patients. However, before establishing such a claim, it is worthwhile to see if there are alternative models of the data that are equally predictive but do not require including such tests.

We present the following definition to facilitate the discussion learning rules.

Definition 1 (Globally Predictive Test)

A test is globally predictive of Class_i iff $P(\text{Class}_i|\text{Test}) > P(\text{Class}_i)$

Definition 2 (Locally Predictive Test)

A test is locally predictive of Class_i in a Context iff $P(\text{Class}_i|\text{Test}\&\text{Context}) > P(\text{Class}_i|\text{Context})$ where Context is some Boolean combination of tests.

In this paper, we explore the implications of biasing rule learners to avoid using tests that are locally predictive of class memberships but are not globally predictive. A single rule that predicts class membership as a conjunction of globally predictive tests is an example of a simple causal schema: multiple necessary causes [12]. A set of such rules that enumerate alternative means of predicting class membership represents another simple causal schema: multiple sufficient causes. However, a rule that uses a test that is locally but not globally predictive is evoking a more complex causal schema in which there is an interaction among the variables. A predictive relationship involving such an interaction among variables is more difficult for people to learn from data [13]. We argue that to match the cognitive bias of human learners, knowledge discovery systems should avoid creating rules with locally predictive tests that are not globally predictive unless such tests are truly necessary to increase the accuracy of this model.

2 Background: Rule Learners

In this work, we will extend a rule learning system to implement the globally predictive bias. We will use FOCL [6] as a representative of this family of algorithms. FOCL is derived from Quinlan's [14] FOIL system. FOIL is designed to learn a set of rules that distinguish positive examples of a concept from negative examples.

FOIL operates by trying to find a rule that is true of as many positive examples as possible and no negative examples. It then removes the positive examples explained by that rule from consideration and finds another rule to account for other positive examples. It repeats this rule learning process until all of the positive examples are explained by some rule. Each rule can be viewed as a description of some subgroup of examples.

To learn an individual rule, FOIL first considers all possible rules consisting of a single test. It selects the best of these according to an information-gain heuristic that favors a test that is true of many positive examples and few negative examples. Next, FOIL specializes the rule using the same search procedure and information-based heuristic, considering how conjoining a test to the current rule would improve it by excluding many negative examples and few positives. This specialization process continues until the rule is not true of any negative examples, resulting in a single rule that is a conjunction of tests.

FOCL follows the same procedure as FOIL to learn a set of rules. However, it learns a set of rules for each class (such as low risk and high) enabling it to also deal with problems that have more than two classes. The rule learning algorithm is run once for each class, treating the examples of that class as positive examples and the examples of all other classes as negative examples. This results in a set of rules for each class. In this paper, we restrict our attention to a simple but effective procedure for converting a set of rules for each class into a single decision list such as that shown in Table 1. The learned rules are ordered by the Laplace estimate of the rules' accuracy [2] and the most frequent class is used as a default class.

When determining which test to add to the rule, FOCL (as well as other rule learners) considers tests in the context of the previous rules that were learned. The examples used to determine which test is best are those that are not true of any rule body that was learned previously and those that are true of the previous tests in the current rule. As a consequence, for all but the first test of the first rule, this family of algorithms can select a test that is locally predictive but not globally predictive. In the next section, we consider biasing rule learners to consider both the global and local predictability.

3 The Globally Predictive Test Bias

We experiment with two forms of the globally predictive test bias: a restriction bias and a preference bias. For the restriction bias, the procedure for selecting the best test is modified to exclude a test from consideration when learning a rule for $Class_i$ unless $P(Class_i|Test) > P(Class_i)$. The restriction bias therefore selects the globally predictive test that is best in the local context to add to a clause under consideration.

The preference bias prefers tests that are globally predictive. It will select a test that is not globally predictive if it is significantly better than the best locally predictive test that is globally predictive. First, the best test in the local context is found. If it is globally predictive, it is used in the rule. If it is not, the best test in the local context that is globally predictive is found. The two tests are then compared. If the globally predictive test is a significantly worse predictor in the local context than the test that is not globally predictive, the test that is not globally predictive is used in the rule. Otherwise, the test that is globally predictive is used. A χ^2 test is used to determine if there is a significant difference between the two tests. By default, if the probability that the two tests differ is greater than 0.75, then the locally but not globally predictive test is used. In our experiments, we determine the value of this probability parameter using cross-validation.

Whether the globally predictive test bias is useful in some domains is an empirical question. Clearly, the ability to have a locally but not globally predictive test will be useful in some problems such as those in which there are interactions among variables. However, this additional degree of freedom may be harmful in other domains resulting in inaccurate or confusing rules. In the next session, we investigate experimentally whether the globally predictive test restriction bias is useful.

3.1 Experiment 1: Restriction Bias

Here we report the results of running experiments on 16 problems selected from industrial and medical research projects at UCI and the UCI Repository of Machine Learning Databases [15]. Most of the data sets are available on the Internet at the UCI archive. The following additional data sets are used:

- Admissions: This database contains data on 312 high school students that were admitted to UC Irvine and the class label indicates whether the students enrolled at UCI. The attributes are scores on standardized tests and descriptive information such as age, gender, and residency.
- CERAD: Data collected by the Consortium to Establish a Registry for Alzheimer's Disease (CERAD). The particular problem of interest is to identify patients with early signs of dementia. The database contains 315 examples with the dementia status of each patient and the results of two commonly used cognitive tests for dementia screening, the Blessed Orientation, Memory and Concentration test and the Mini-Mental Status.
- FAQ: This is also data on screening for early signs of dementia collected by the UCI Center for Brain and Aging. The database contains 347 examples with the dementia status of each patient and the results of the Functional Activities Questionnaire.
- Retardation: This contains 4302 examples of newborn infants collected by the National Collaborative Perinatal Project of the National Institute of Neurological and Communicative Disorders and Stroke. The task is to determine whether an infant has mild mental retardation.
- Staging. This problem is to determine the severity of dementia of a patient from results of cognitive and neuropsychological tests. It contains 765 patient records.

We conducted an experiment in which we compared FOCL with the global predictive test restrictive bias to FOCL without this bias. The goal was to determine whether this bias is useful in practice. For each domain, we used paired ten-fold cross-validation of FOCL with and without this bias and computed the average accuracy for each of the databases. Table 2 also lists the average accuracy with and without the bias. We performed a paired t-test to determine whether there is a significant difference in using the bias on each data set. Figure 1 shows the difference in accuracy between using FOCL with the bias and using FOCL without the bias. Those problems in which a significant difference was found at the .05 level or greater are shown in black.

The results demonstrate that this bias results in a significant increase in accuracy on three data sets and a significant reduction on one. This shows that there are situations in which the extra freedom allowed by selecting a test that is locally predictive but not globally predictive is harmful. There are also situations in which the globally predictive test bias is harmful. The King-Rook-King-Pawn is an example of where the bias would not be expected to work well. This is a chess problem where the goal is to determine whether the white player with a king and rook can defeat a black player with a king and a pawn. The attributes in this problem correspond to features describing the locations of the pieces (e.g., the white king is in the last row). In this problem, it is the interaction among several features that determines whether white can win.

The CERAD data set is a particularly interesting illustration of the power of this bias. The attributes represent replies to questions designed to assess cognitive capabilities and those tests that are globally predictive of dementia represent incorrect answers to the questions. Those tests that are locally but not globally predictive of dementia are correct answers to questions. Although on a subsample of data they appear to be predictive of dementia, this is not a very reliable pattern when tested on unseen data.

Table 2: Databases used in the experiments and results of Experiment 1.

Problem	**Classes**	**Without Bias**	**Restriction Bias**
admissions	2	.696	.711
bupa	2	.664	.716 *
CERAD	2	.917	.949 *
colic	2	.827	.830
FAQ	2	.865	.870
glass	7	.674	.683
hepatitis	2	.800	.807
ion	2	.829	.838
krkp	2	.989	.971 *
mushrooms	2	.998	.999
pima	2	.724	.758 *
retardation	2	.701	.691
staging	4	.666	.671
voting	2	.936	.945
wine	3	.944	.950
wisc	2	.681	.705

Furthermore, rules that indicate that getting a question correct is a sign of dementia are puzzling to the experts in the domain. Others (e.g., Holte, Acker, & Porter [16], Pagallo & Haussler [7], Murphy & Pazzani [17], Vilalta, Blix, & Rendell [18]) have also reported on the problems associated with unreliably estimating descriptive statistics from small groups of examples and have proposed solutions based upon preventing examples from being partitioned into small groups. Here, we explore a different approach in which we reduce the hypothesis space to mitigate this problem. While the prior work has focused on improving the accuracy, we are motivated by improving the understandability of learned rules without reducing the accuracy.

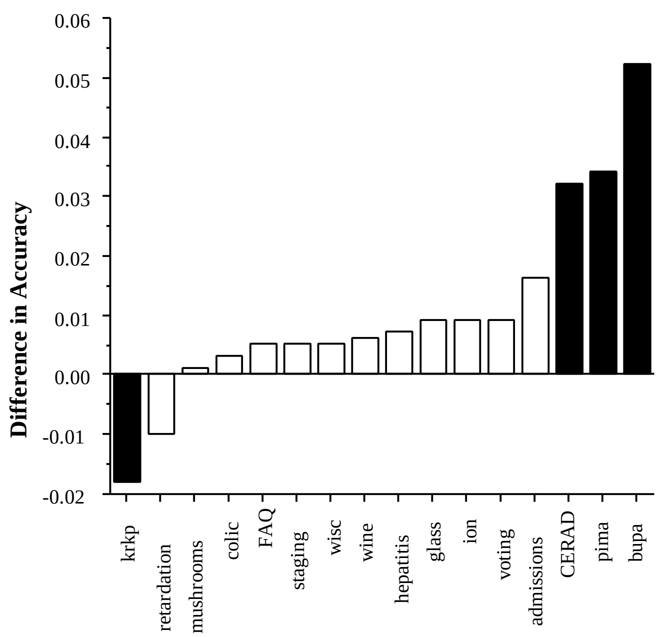

Fig. 1. Difference in accuracy between FOCL with the restriction bias and FOCL without this bias. Significant differences are shown in black. Positive values indicate that more accurate results are obtained when using the bias.

The above discussion suggests that the globally predictive bias may aid in preventing overfitting. By requiring that tests be both globally and locally predictive, some unreliable tests may be eliminated from consideration. The previous experiment did not use any pruning method to reduce the effects of overfitting. In the next section, we report on an experiment designed to determine whether the globally predictive bias provides additional benefits when pruning.

3.2 Experiment 2: Restriction Bias with Pruning

The methodology used in this experiment is identical to the methodology used in Experiment 1. The only change is that reduced error pruning is used with FOCL both with and without the global predictive test restriction bias. Brunk & Pazzani [19] showed that reduced error pruning was more effective than the minimum description length heuristic used by FOIL on a variety of problems. Reduced error pruning operates by dividing the training data into two partitions. One partition (70% of the training data) is used to learn rules. The remaining training data is used in pruning. Two operators are used in pruning: deleting a rule and deleting the last test of a rule. These operators are applied to each rule learned and if a change improves the accuracy of the rules (as estimated on the pruning set), the change that results in the largest increase in accuracy is made permanent. Pruning is repeated until no change increases the accuracy of the learned rules. Figure 2 shows the difference in accuracy when reduced error pruning is used between FOCL with the global predictive test bias and FOCL without this bias. The black bars show significant differences in accuracy and positive values indicate that the bias was beneficial on that domain.

The results of these experiments indicate that the global predictive test bias provides an additional benefit over pruning. In 4 of the 16 problems, there is a significant increase in accuracy, while on two problems, there is a significant decrease. Furthermore, many of the increases in accuracy are greater than the largest decrease in accuracy. One possible reason that the bias provides benefits even when pruning is that pruning deletes tests that are unreliable but doesn't allow for the replacement of these tests with more reliable tests. In contrast, if the bias eliminates a test that is locally but not globally predictive, it finds a new test that is both globally and locally predictive to explain the data. Of course, this test is then also subject to the same pruning algorithm and is only retained if it is needed to increase accuracy on the pruning set.

Although the bias is useful on many problems, there are still some problems in which the bias is harmful. We would expect such a result with any bias for theoretical reasons (cf. Schaffer [20]) and with this particular bias we'd expect it to have problems when there are interactions among variables that make some variables locally but not globally predictive of class membership. In the next experiment, we relax the bias by preferring tests that are globally predictive.

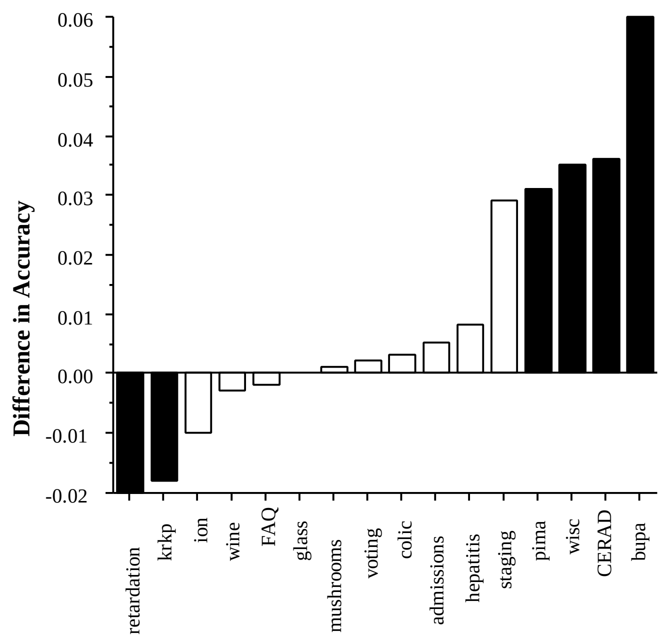

Fig. 2. Difference in accuracy between FOCL with the restriction bias and FOCL without this bias when using reduced error pruning.

3.3. Experiment 3: The globally predictive test preference bias

The globally predictive test bias is too restrictive for some domains. In this section, we explore a related preference bias. The preference will select a test that is not globally predictive if it is significantly better than the best locally predictive test that is globally predictive. In the experiments, a χ^2 test will be used to determine if there is a difference between the two tests. We use 5-fold cross validation to determine the best setting for the probability that there is a difference selecting from 0.05, 0.25, 0.5, 0.75 and 0.95. The experiment below is run using the same methodology as the previous two experiments. On each trial, the threshold for the χ^2 test is found by cross-validation on the training data before the global predictive test bias is compared to the accuracy of FOCL with this preference bias. The average difference in accuracy is plotted in Figure 3 for the 16 domains.

The results graphed in Figure 3 show that there are 6 domains in which the preference bias provides a significant increase in accuracy. Although there are decreases in accuracy, these are all less than one percent and none of these are significant. This suggests that the cross-validation test is generally effective at determining how large a difference is needed between the best locally but not globally predictive test and the best locally predictive test that is globally predictive to ignore the influence of the global predictiveness of a test.

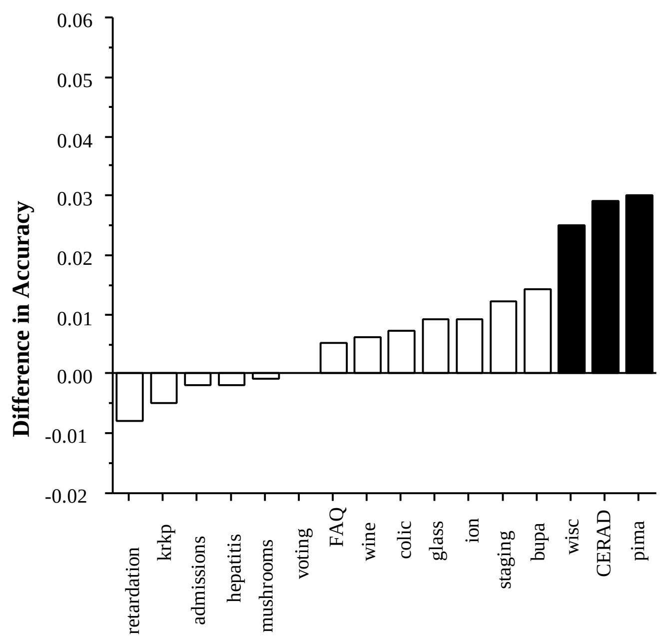

Fig. 3. Difference in accuracy between FOCL with the preference bias and FOCL without this bias.

An advantage of the preference bias over the restriction bias is that the preference bias does learn rules with tests that are locally predictive but not globally predictive. Such tests may represent important insights to convey to domain experts. However, unlike a system without any bias for globally predictive tests, the preference bias first ensures that there is not another alternative that is globally predictive. As a consequence, it includes fewer such tests in the rule, making it easier for an expert to verify that a useful interaction among variables has been found.

4 Discussion

In previous work (Pazzani, Mani & Shankle [21]), we addressed the problem of learning algorithms including counterintuitive tests in rules by having an expert provide "monotonicity constraints". For nominal variables, a monotonicity constraint is expert knowledge that indicates that a particular value makes class membership more likely. For numeric variables, a monotonicity constraint indicates whether increasing or decreasing the value of the variable makes class membership more likely. Lee, Buchanan, & Aronis [22] introduce similar expert constraints to the RL rule learner to make carcinogenicity more understandable and more accurate. Here, we show that much of the same effect could be achieved without consulting an expert by considering the global predictiveness of the training data. One advantage of the current approach is that it doesn't require an expert and may be applied more easily to many databases.

The expert monotonicity constraint bias was applied to the CERAD database. Pazzani, Mani & Shankle [21] report an accuracy of 90.7% using this constraint and 90.6% without. In contrast, C4.5 was 86.7% accurate, C4.5 rules was 82.6% accurate and a naïve Bayesian classifier was 91.2%. The globally predictive test restriction bias obtained an accuracy of 94.4% on this database, substantially higher than the monotonicity constraint bias. There are two reasons for this difference in accuracy. First, one monotonicity constraint for a nominal value did not turn out to be globally predictive. This test was ignored when using monotonicity constraints but is frequently used with the global predictive test bias. Second, monotonicity constraints are not as specific as the global predictive test bias for numeric variables. In particular, a test includes both a comparison (such as greater than) and a specific numeric threshold. The global predictive test bias determines whether a test is globally predictive while the monotonicity constraint represents more general information about whether increasing values tend to make the class more likely. As a consequence, when using monotonicity constraints it is possible to have tests on numeric values that are locally but not globally predictive.

The globally predictive test bias represents a form of simplicity bias. However, in this case simplicity is not a syntactic property of the representation. Rather, it is a preference for a simple causal mechanism in which the influence of a variable on an outcome is not inverted in the context of other variables. That this bias is effective in increasing the accuracy of learned models is evidence that the databases commonly collected have such simple causal models. Similarly, the success of the bias may help to explain why replacing greedy searches for rules with more exhaustive searches (e.g., Rymon [23]; Webb [24]) has not been beneficial on most databases. Additional search would be useful to detect complex interactions among variables to find sets of locally but not globally predictive tests. However, if such situations are uncommon, the additional search is likely to overfit the data [25].

The globally predictive test bias, especially the restriction, could be viewed as a form of feature selection. For example, some approaches order variables by informativeness, a global criterion and select only the most informative prediction [26]. There are several differences however. First, we are selecting tests, not just variables. Second, we make this decision separately for each class. The most significant difference comes in the preference bias when we are favoring tests that meet these criteria but do not eliminate any test from consideration.

The original motivation of this work has been to improve expert acceptance of the results of knowledge discovery in databases. Experiments are in progress in which experts and novices judge the plausibility of rules learned with and without these global predictive constraints. Earlier experiments showed that experts preferred rules that obeyed monotonicity constraints and given the close relationship between monotonicity constraints and the global predictive test bias we are hopeful that the bias will prove useful in making the results of KDD more acceptable to experts.

5 Conclusions

We have explored the implications of biasing rule learners to create tests that are both globally and locally predictive of class membership. The results show that this bias improves the accuracy of learned models on a variety of domains. The knowledge discovery process is often viewed as an iterative process of modeling data with learning algorithms and changing the representation of the data or the parameters of the algorithm in an attempt to gain insight from the data. The global predictive test bias represents another tool in the toolkit that is intended to avoid overly complex models when simpler explanations of the data are possible.

Acknowledgements

We thank William Cohen for running RIPPER on synthetic data we developed to illustrate inverted influences. Comments by Pedro Domingos, Dennis Kibler, Cathy Kick, Richard Lathrop, Subramani Mani and the graduate students at UCI were helpful in clarifying the ideas presented in this paper.

References

1. Michalski, R., Mozetic, I., Hong, J., & Lavrac, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. *Proceedings of the 5th National Conference on Artificial Intelligence.* Philadelphia, PA: Morgan Kaufmann. 1041-1047.
2. Clark, P. & Niblett, T. (1989). The CN2 Induction Algorithm*Machine Learning,3*, 261-284.
3. Goodman, R., & Smyth, P. (1989). The induction of probabilistic rule sets: the ITRULE algorithm, *Proceedings of the Sixth International Machine Learning Workshop*, (pp. 129-132). Los Altos, CA: Morgan Kaufmann.
4. Quinlan, J.R. (1992). *C4.5: Programs for Machine Learning.* Los Altos, CA:Morgan Kaufmann.
5. Quinlan, J.R. (1990). Learning logical definitions from relations. *Machine Learning,* 5, 239–266.
6. Pazzani, M., & Kibler, D. (1992). The utility of knowledge in inductive learning. *Machine Learning,* 9, 57-94.
7. Pagallo, G., & Haussler, D. (1990). Boolean feature discovery in empirical learning.
8. Cohen, W. (1995). Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, Lake Tahoe, California.
9. Rivest, R. (1987). Learning decision lists. *Machine Learning,* 2:229 - 246.
10. Ali, K., & Pazzani, M. (1993). HYDRA: A noise-tolerant relational concept learning algorithm. *The International Joint Conference on Artificial Intelligence,* Chambery, France
11. Mani, M., McDermott, S., & Pazzani, M. (1997). Generating Models of Mental Retardation from Data with Machine Learning. *Proceedings IEEE Knowledge and Data Engineering Exchange Workshop* (KDEX-97), p. 114-119, IEEE Computer Society.
12. Kelley, H. (1983). The process of causal attribution. *American Psychologist*, 107-128.
13. Pazzani, M., & Silverstein, G. (1990). Feature selection and hypothesis selection: Models of induction. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, (pp. 221-228). Cambridge, MA: Lawrence Erlbaum.

14. Quinlan, J.R. (1990). Learning logical definitions from relations. *Machine Learning*, 5, 239–266.
15. Merz, C.J., & Murphy, P.M. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science. *Machine Learning* 5(1):71-100.
16. Holte, R. Acker, L. and Porter, B. (1989). Concept learning and the problem of small disjuncts. *Proceedings International Joint Conference on Artificial Intelligence*. pp. 813-818.
17. Murphy, P., & Pazzani, M. (1991). ID2-of-3: Constructive induction of *m-of-n* discriminators for decision trees. *Proceedings of the Eighth International Workshop on Machine Learning* (pp. 183-187). Evanston, IL:Morgan Kaufmann.
18. Vilalta, R., Blix, G. and Rendell, L. (1997). Global Data Analysis and the Fragmentation problem in Decision Tree Induction 9th European Conference on Machine Learning *Lecture Notes in Artificial Intelligence*, Vol. XXX. Springer-Verlag, Heinderberg, pp 312-326. *Workshop on Machine Learning* (pp. 183-187). Evanston, IL: Morgan Kaufmann.
19. Brunk, C., & Pazzani, M. (1991). An investigation of noise tolerant relational learning algorithms. *Proceedings of the Eighth International Workshop on Machine Learning* (pp. 389-391). Evanston, IL: Morgan Kaufmann.
20. Schaffer, C. (1994). A conservation law for generalization *Proceedings of the 11th International Conference of Machine Learning,* New Brunswick. Morgan Kaufmann.
21. Pazzani, M., Mani, S., & Shankle, W. R. (1997). Comprehensible knowledge-discovery in databases. In M. G. Shafto and P. Langley (Ed.) *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, pp. 596-601. Lawrence Erlbaum.
22. Lee, Y. Buchanan, B., & Aronis, J. (in press). Knowledge-Based Learning in Exploratory Science: Learning Rules to Predict Rodent Carcinogenicity. *Machine Learning*.
23. Rymon, R. (1993). An SE-based characterization of the induction problem. *Proceedings of the 10th International Conference of Machine Learning,* (pp. 268-275). Amherst, MA: Morgan Kaufmann.
24. Webb. G. (1993) Systematic search for categorical attribute-value data-driven machine learning. *Proceedings of the Sixth Australian Joint Conference on Artificial Intelligence*, (pp. 342-347). Melbourne: World Scientific.
25. Quinlan, J. R., & Cameron-Jones, R. (1995). Oversearching and layered search in empirical learning. *Proceedings Fourteenth International Joint Conference on Artificial Intelligence*, (pp. 1019-24). Morgan Kaufmann, Montreal.
26. Chen, M., Han, J., and Yu, P. (1996) Data Mining: An Overview from a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6.