Exploring Alternative Biases Prior to Learning in Scientific Domains

Bruce G. Buchanan and **Yongwon Lee**
Intelligent Systems Laboratory
University of Pittsburgh
Pittsburgh, PA 15260

Abstract

Before machine learning can be applied to a new scientific domain, considerable attention must be given to finding appropriate ways to characterize the learning problem. A central idea guiding our work is that we must make explicit more of the elements of a program's bias and understand the criteria by which we prefer one bias over another. We illustrate this exploration with a problem to which we have applied the RL induction program, the problem of predicting whether or not a chemical is a likely carcinogen.

Introduction

Inductive reasoning over a collection of scientific data is not a single event, because considerable exploration is required to characterize the data and the target concepts appropriately. Mitchell (Mitchell 1980), Provost (Provost 1992), and others refer to this process as searching a bias space for an inductive learning program, where the bias is everything in the program that influences learning other than the positive and negative examples used for learning. In the ongoing work described here, we are examining issues involved in the preliminary characterization of a learning problem.

A central idea guiding our work is that we must make explicit more of the elements of a program's bias and understand the criteria by which we prefer one bias over another. In automating the search of a bias space, we can separate out a class of pragmatic considerations, called here an induction "policy", that can be seen to guide the selection of an appropriate bias.

A simple model of induction is all that is required to illustrate this central idea, and is the root of most procedures used by induction programs: First, collect preclassified positive and negative examples of a target concept. Second, find characteristics shared by all the positive examples and none of the negatives. Finally, assert the conjunction of shared features as a definition.

This procedure, which can be implemented as simple set intersection, is not an adequate procedure for real-world science, however. The primary reasons why it breaks down are first, a scientist must first characterize which target concept(s) are important and label the examples as positive or negative with respect to the target(s), and second, he or she must characterize the data with respect to features that are likely to yield meaningful associations with the target phenomena. Moreover, empirical data are subject to many shortcomings. The data may be:

incomplete – descriptions of objects or events are missing values of some attributes; there are too few data points for significant inferences

redundant – not all attributes are independent; some objects or events may be described several times

noisy – some measurements are erroneous because of random "white noise"

erroneous – some observations and measurements are subject to random or systematic errors, *e.g.* reporting and transcription errors and calibration errors

voluminous – extremely large collections of data can be sampled, but the volume of data can be difficult to manage.

The RL Program

We have been working with an induction program, called RL (Clearwater & Provost 1990; Provost *et al.* 1993), to examine extensions to the basic model that will allow mechanized induction to be useful to scientists. We briefly describe work in one domain, finding rules that can be used to predict likely carcinogenic activity of chemicals. Here we emphasize some of the exploratory work that we performed (which was only partly automated) in order to set up the conditions under which RL could find interesting rules.

RL is a heuristic search program. Using a top-down generator of successively more specialized rules and many semantic and syntactic criteria of rule "goodness", RL searches a combinatorial space of conjunctive rules to find a universally quantified disjunction of rules that can adequately cover, or explain, the training data. Each rule is a conditional sentence (LHS $\Longrightarrow$

RHS), with an implicit universal quantification, with the LHS a conjunction of features and the RHS the named target concept (K):

$$f_1 \wedge \ldots \wedge f_n \Longrightarrow K(x).$$

This is interpreted to mean that any object or event that is described as having features, $f_1, \ldots, f_n$, is a member of the target class K. A feature is a predicate relating the value of an attribute with a constant, *e.g.*, the molecular weight of a chemical is greater than 200. For example, a disjunction of two rules may assert

$$\begin{aligned} \forall x \quad & [(A_1(x) = v_1 \wedge A_2(x) > v_2 \wedge A_3(x) = v_3) \\ & \quad \vee \\ & (A_1(x) = v_1 1 \wedge A_4(x) < v_4) \\ & \\ & \Longrightarrow K(x)] \end{aligned}$$

At each step, the generator either adds a new attribute name to the LHS of a rule or associates a more specialized value with an existing attribute. For attributes whose values are taken from a continuous range, RL automatically finds discrete intervals that have predictive power. For large sets of symbolic values, RL can use a pre-defined hierarchy of values to make its search more efficient and make the resulting rules more general.

RL's bias includes the vocabulary used to name and define features, syntactic constraints on the number of conjuncts in a single rule or the number of disjuncts in a rule set, and thresholds that define sufficient empirical support needed by a rule or rule set. Dependencies among the attributes, and their ranked importance, can be specified, as well as relationships that the program should take as given. The parts of RL's bias that have been made explicit and easily changed we have referred to in prior publications as the partial domain theory.

The dilemma of applying RL, or any induction program, to problems of real science, is that the "correct" bias is not always known in advance. To make matters worse, a bias that is "good enough" is partly recognized through RL's ability to find "good" rules, where the criteria of goodness are part of the bias. The considerations that guide the selection of a suitable bias we have termed an "induction policy" (Provost & Buchanan 1994).

The Problem of Predicting Chemical Carcinogens

We have applied RL to several different data sets, including data collected by the National Toxicology Program on the propensity of many chemicals to cause cancer (Lee *et al.* 1994). Because many human lives are at stake, this is an important problem; also, long-term epidemiological studies on people or long-term studies with rodents may each cost several millions of dollars. Rather than attempt *a priori* predictions of biological activity from chemical structure, our collaborator[1] asked us to investigate the use of data from inexpensive, short-term biological assays performed on bacteria or other cells cultured in the laboratory. An important part of "the problem" is to find an adequate descriptive vocabulary within which meaningful relationships can be found.

It is not universally acknowledged that short-term assays carry enough information to be useful. For example,

> "... It is clear that even with a battery of assays, not all rodent carcinogens are *in vitro* mutagens, nor are all *in vitro* mutagens rodent carcinogens. If current in vitro short-term assays are expected to replace long-term rodent studies for the identification of chemical carcinogens, then that expectation should be abandoned." (Tennant *et al.* 1987)

Instead of relying entirely on data from such assays, we chose to include some other easily-obtained information as well.

Ambiguities and likely errors in the data also present a difficulty to be overcome before running a machine learning program. A panel of experts in the National Toxicology Program has made collective judgments about the carcinogenicity of the chemicals in the database based on long-term studies of rodents. Any of about two dozen organs in male or female mice or rats may show tumors or pre-cancerous lesions after a year's exposure (or 3, 6, or 9 months' exposure) to various doses of a chemical. From these data the panel judged that a chemical is (NC) definitely not carcinogenic, (CA) carcinogenic in rodents because of lesions in more than one organ of male and female mice and rats, or in a gray area between these two. The test results themselves are not entirely reproducible across laboratories, for example, because human judgment is involved in categorizing cells seen under a microscope.

One short-term assay with *Salmonella* has been found to have about 60% predictive accuracy by itself when the test is positive. In one study we chose to focus on the problem cases when the *Salmonella* test is negative (written as "SAL-"). In our training data of 135 SAL- chemicals, 51.1% (69) were coded as likely carcinogenic (CA) and 48.9% (66) as definitely not carcinogenic (NC), so a SAL- result provides no information. A test set of 24 SAL- chemicals was used for evaluation.

As our initial vocabulary we started with eleven attributes of chemicals, grouped in three sets as shown below. All are relatively easy, quick, and inexpensive to measure or calculate.

[1] Prof. Herbert Rosenkranz, Dept. of Environmental and Occupational Health, Univ. of Pittsburgh

Five short-term assays (symbolic)	
SAL	*Salmonella* mutagenicity
SCE	Sister chromatid exchanges in Chinese hamster ovary cells
ChrA	Chromosomal aberrations in cultured cells
Tran	Induction of cell transformation in cultured cells
Cyto	Cell toxicity in cultured 3T3c cells
Four physical chemical properties (continuous)	
ELN	Electron negativity
LgP	Log of octanol-water partition coefficient
Mwt	Molecular weight
Wsolub	Water Solubility
Two Mtd values (continuous)	
Mmtd	Maximum tolerated dose for mice
Rmtd	Maximum tolerated dose for rats

Some of the questions we asked were, How adequate is this vocabulary?, Which attributes make the most difference?, Can a subset predict as accurately as the full set?, and Which subset should be used?

Results

Many experiments (ten-fold cross validations) were performed to answer these questions. In each experiment, RL was provided with a different set of vocabularies (attributes), criteria for determining "good" rule, and/or different policy for weighing the evidence that a rule provides. Four performance statistics were determined in each experiment through averaging the results of ten-fold cross-validation with the 135 chemicals: (1) sensitivity ("Sens") for predicting rodent carcinogens, (2) specificity ("Spec") for predicting rodent non-carcinogens, (3) overall accuracy ("Acc") for predicting rodent carcinogenicity of chemicals and (4) the frequency with which predictions were made at all ("Pred"). Base-line statistics were first determined for subsets of the last six attributes, without the short-term assays. These are shown in Table 1.

Then, as shown in Table 2, many more experiments were run adding combinations of the short-term assays to the six attributes used to establish the baseline.

Since several rules in a disjunctive set may match a new case, and some of their predictions may be inconsistent, a policy is needed on how to weigh the evidence and make a single prediction. We performed many more experiments over the training data with six different strategies for weighing evidence:

- use the conclusion of the first rule whose LHS matches
- use the conclusion of the strongest rule[2].
- use the conclusion of the majority of rules
- use the conclusion of the plurality of rules
- use the conclusion of weighted voting[2]
- use the conclusion of MYCIN's CF combination function[2]

Running many similar experiments with other aspects of the bias, we determined the vocabulary and conditions under which RL appeared to find meaningful rules from the training data. Using such vocabulary and conditions (other biases and the policy of weighing rule evidence), a set of rules was learned from all 135 training chemicals. We applied the rule set to predict carcinogenicity of a set of 24 SAL-chemicals which were not included in the set of 135 chemicals used in this study. These 24 are among the 44 chemicals (Tennant *et al.* 1990) which formed the subject of the International Workshop of Predicting Chemical Carcinogenesis in Rodents (Parry 1994; Ashby & Tennant 1994). Results are summarized in Table 3 and also compared to experts' decisions (Tennant *et al.* 1990).

An example of one rule in the set R3 learned by RL is shown below. The SAL- feature is implicit in all the rules but is shown explicitly here as a reminder.

IF {SAL-} and
(LgP > 3.325) and
(ChrA negative)
THEN Carcinogenic

In the training set, this rule covered 20 carcinogens correctly and 2 non-carcinogens incorrectly. In the test set of 24 new chemicals, the rule covered 4 chemicals, 3 of 8 carcinogens and 1 of 7 non-carcinogens, and none of the chemicals whose carcinogenicity was equivocal or unknown were covered.

Our policy guiding the selection of a rule set included both accuracy ("Acc") and coverage ("Pred"). One reason that we selected attributes ChrA and SCE instead of other combinations from the experiments partially reported in Table 2 was was that there were few missing values of these attributes. (More specifically, since rules named only one or the other of these two short-term assays in their LHS's, coverage was dependent on the number of missing values of either one or the other attribute.) The cost of failing to make a prediction is not as great as the cost of making a false prediction, but it is not negligible in this domain.

Conclusion

The present research has shown the importance of empirical explorations of alternative biases in one urgent problem of science. In related work, Provost (Provost 1992) has shown how some of these considerations can be used to guide an automated search of the bias space. We believe machine learning programs must have the flexibility to work with alternative biases and that the considerations guiding the selection of a bias need to be made more explicit.

[2]The weight, or CF, of a single rule is its predictive accuracy

Table 1: Results (in %) of seven ten-fold cross validation experiments without short-term biological assays.

Exp#	Info used in rules	Sens	Spec	Acc	Pred
1	ELN Mwt Wsolub LgP Mmtd Rmtd	54.5	64.6	57.5	96.3
2	Mwt Wsolub LgP Mmtd Rmtd	49.2	74.4	60.1	90.0
3	ELN Wsolub LgP Mmtd Rmtd	56.1	60.2	56.1	83.7
4	ELN Mwt LgP Mmtd Rmtd	47.8	63.3	53.2	88.2
5	ELN Mwt Wsolub Mmtd Rmtd	53.0	65.2	56.6	83.4
6	ELN Mwt Wsolub LgP Rmtd	58.1	64.2	59.0	89.0
7	ELN Mwt Wsolub LgP Mmtd	55.4	64.5	57.8	95.6

Table 2: Selected results (in %) of seven ten-fold cross validation experiments adding different combinations of short-term biological assays.

Exp#	short-term assays used in addition to 4 phy/chem features and 2 Mtds	Sens	Spec	Acc	Pred
1	no short-term assay (Table 1)	54.5	64.6	57.5	96.3
2	ChrA	78.1	64.5	69.2	95.6
3	SCE	76.8	59.4	66.8	95.6
4	ChrA SCE	74.6	61.4	66.1	95.5
5	ChrA SCE Cyto	64.3	69.8	64.7	95.6
6	ChrA SCE Tran	62.4	68.4	63.4	96.3
7	ChrA SCE Tran Cyto	60.7	69.5	63.4	96.3

Table 3: Results of Testing on the test set of 24 new chemicals, consisting of 8 rodent carcinogens, 7 rodent non-carcinogens, 3 whose carcinogenicity was determined equivocal, and 6 chemicals whose carcinogenicity was not known. R3 predicted all 8 carcinogens and 4 of 7 non-carcinogens. R3 also made predictions on all 3 equivocals (of which 2 are identical to the experts' decisions) and 5 of 6 unknowns (of which 4 are equal to the experts' decisions). Thus, overall, R3 made 8 predictions on these 9 chemicals, with 6 of them equal to what experts predicted using a much larger amount of information than was available to RL.

	Sens	Spec	Acc
R3 (25 rules–13 for CA and 12 for NC)	100.0 (=8/8)	57.1 (=4/7)	80.0 (=12/15)
Experts' decisions (Tennant *et al.* 1990)	75.0 (=6/8)	85.7 (=6/7)	80.0 (=12/15)

Acknowledgements

We gratefully acknowledge financial support, in part, from the W.M. Keck Foundation, the Department of Defense (DAAA21-93-c-0046), and the Center of Alternatives to Animal Testing. We thank Dr. Herbert Rosenkranz, who has been most generous with his time in many discussions of this problem, and Dr. Foster Provost, whose work on RL and on searching a bias space provided many valuable insights.

References

Ashby, J., and Tennant, R. 1994. Prediction of rodent carcinogenicity for 44 chemicals: Results. *Mutagenesis* 9:7–15.

Clearwater, S., and Provost, F. 1990. RL4: A tool for knowledge-based induction. In *Proceedings of Tools for Artificial Intelligence 90*, 24–30. IEEE Computer Society Press.

Lee, Y.; Rosenkranz, H.; Buchanan, B.; Mattison, D.; and Klopman, G. 1994. Learning rules to predict carcinogenicity of nongenotoxic chemicals in rodents. Technical Report ISL-94-5, Intelligent Systems Laboratory, University of Pittsburgh. To be appeared in Mutation Research.

Mitchell, T. 1980. The need for biases in learning generalizations. Technical Report CBM-TR-117, Department of Computer Science, Rutgers University.

Parry, J. 1994. Detecting and predicting the activity of rodent carcinogens. *Mutagenesis* 9:3–5.

Provost, F., and Buchanan, B. 1994. Inductive policy: The pragmatics of bias selection. Technical Report ISL-94-4, Intelligent Systems Laboratory, University of Pittsburgh. Submitted to Machine Learning.

Provost, F.; Buchanan, B.; Clearwater, S.; Lee, Y.; and Leng, B. 1993. Machine learning in the service of exploratory science and engineering: A case study of the RL induction program. Technical Report ISL-93-6, Intelligent Systems Laboratory, University of Pittsburgh.

Provost, F. 1992. *Policies for the Selection of Bias*

in Inductive Machine Learning. Ph.D. Dissertation, Computer Sceince Department, University of Pittsburgh. No. 92-34.

Tennant, R.; Margolin, B.; Shelby, M.; Zeiger, E.; Haseman, J.; Spalding, J.; Caspary, W.; Resnick, M.; Stasiewicz, S.; Anderson, B.; and Minor, R. 1987. Prediction of chemical carcinogenicity in rodents from in vitro genetic toxicity assay. *Science* 236:933–941.

Tennant, R.; Spalding, J.; Stasiewicz, S.; and Ashby, J. 1990. Prediction of the outcome of rodent carcinogenicity bioassays currently being conducted on 44 chemicals by the national toxicology program. *Mutagenesis* 5:3–14.