Causal Mechanisms and Classification Trees for Predicting Chemical Carcinogens

Louis Anthony Cox, Jr.
Cox Associates, tony@cox-associates.com
503 Franklin Street, Denver, CO, 80218

Abstract

Classification trees, usually used as a nonlinear, nonparametric classification method, can also provide a powerful framework for comparing, assessing, and combining information from different expert systems, by treating their predictions as the independent variables in a classification tree analysis. This paper discusses the applied problem of classifying chemicals as human carcinogens. It shows how classification trees can be used to compare the information provided by ten different carcinogen classification expert systems, construct an improved "hybrid" classification system from them, and identify cost-effective combinations of assays (the inputs to the expert systems) to use in classifying chemicals in future.

1 INTRODUCTION

One of the most difficult applications challenges for statistical and AI classification technology has turned out to be predicting which chemicals are likely to cause cancer in humans, without performing costly experiments in mice and rats to find out. Part of the difficulty stems from the fact that the term "carcinogen" applies to chemicals that operate by radically different causal mechanisms to produce very different biological responses involving uncontrolled cell proliferation, all of which are referred to as "cancer" (Williams 1996). Learning the concept "carcinogen" from training data therefore requires learning a disjunction of concepts that are heterogeneous in terms of the physical reality being described -- the relevant chemical structures and physiochemical properties, the biological systems affected, and the spectrum of biological responses produced in test systems. The extension of the term "carcinogen" involves an intrinsically complex and heterogeneous ontology that cannot easily be represented by few or simple relations among attributes in a training database.

Despite this complexity, carcinogenicity in mice and rats often predicts carcinogenicity in humans (Ashby and Paton, 1993). More specifically, chemical carcinogens can usefully be subdivided into *genotoxic* carcinogens, which cause cancer by reacting with DNA, and *non-genotoxic* carcinogens, which involve other causal mechanisms leading to stimulated proliferation of Williams 1996, Chevalier 1998). Strongly genotoxic carcinogens often cause cancer in multiple species, sexes, strains, and organs by a common DNA-damaging mechanism (Gold 1991). Therefore, potent mouse- and rat-carcinogens are often considered to be potential human carcinogens. Inferring likely human carcinogenicity for non-genotoxic chemicals, however, is a largely unsolved problem (Ashby 1993). An example of a non-genotoxic mechanism is found in experiments with diesel exhaust (DE), which can cause lung cancer in rats at high, prolonged exposures by forming soot deposits that repeatedly abrade and irritate the lung tissue, eventually depleting protective enzymes and inducing compensating proliferation of cells. This increased proliferation, in turn, raises the probability that at least one cancerous cell will arise. Such non-genotoxic mechanisms tend to be highly species-specific (Ashby 1993, Chevalier 1998). For example, DE does not appear to cause lung cancer or deplete protective enzymes in other species (Cox, 1997).

Genotoxic chemical carcinogens often have structural similarities (such as a "bay region" in a multi-ring organic molecule) that once seemed promising for predicting carcinogenicity. Yet, non-genotoxic carcinogens constitute a miscellany of chemicals, from simple organics like chloroform (Templin 1998,) to complex ones like DE, that increase cell proliferation by various idiosyncratic mechanisms (Williams 1996, Yoshikawa 1996). This creates an inherently deceptive setting for many machine-learning and automated inference or concept-learning programs. Patterns that might prove predictively useful if only genotoxic chemicals were considered become diluted and confounded by non-genotoxic chemicals. The result is that even relatively sophisticated predictive systems often perform poorly when tested on chemicals for which the correct classification is initially unknown (Benigni 97).

This paper introduces a new approach to predicting chemical carcinogens. It is motivated by the observation that different current predictive systems incorporate some complementary and some redundant information about relevant aspects of chemical structures, properties, and effects in various assays and biological systems. Analyzing the empirical performance (i.e., prediction accuracy and failure patterns) of these different

algorithms leads to a relatively rich model of how their errors are interrelated. This, in turn, reveals how their predictions can best be combined to obtain a hybrid model that out-performs any of the individual models that contribute to it.

2 AN ILLUSTRATIVE EXAMPLE

Figure 1 illustrates one such hybrid predictive model, based on the performance data from the ten individual predictive systems summarized in Table 1.

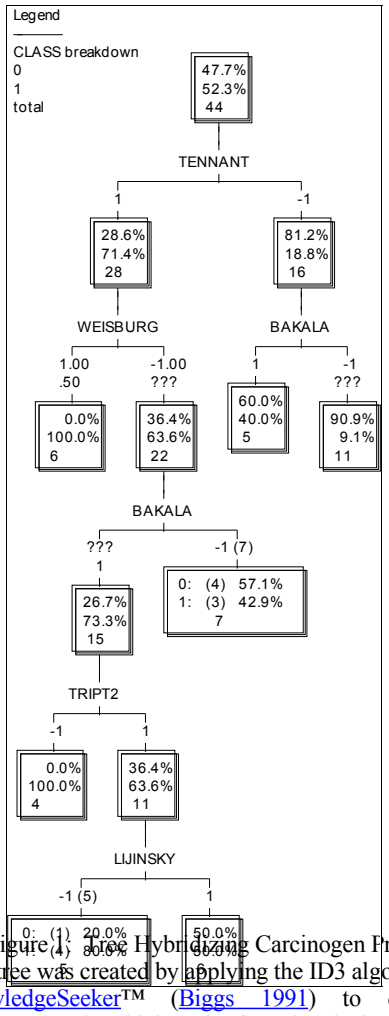

Figure 1. Tree Hybridizing Carcinogen Predictions

This tree was created by applying the ID3 algorithm in KnowledgeSeeker™ (Biggs 1991) to dichotomized outcome data in which each of 44 chemicals was classified as either a rodent carcinogen or as not clearly carcinogenic, based on the outcomes of long-term cancer bioassays in mice and rats. For clarity of exposition, we dichotomized the ternary outcomes used by many prediction systems, which classify chemicals as carcinogens, non-carcinogens, or equivocal/uncertain carcinogens in rodents, e.g., based on whether cancer is predicted to occur in both, neither, or just one of mice and rats. In Figure 1 and subsequent trees, "Class 1" represents a rodent carcinogen in these bioassays, while "Class 0" represents a non-carcinogen or equivocal carcinogen, using Bristol's (1996) summary of bioassay results.

The independent variables entering the classification tree analysis were predictions of carcinogenicity from each of ten individual predictive expert systems (i.e., Tennant, Weisburger, etc.) summarized in Table 1. The predictions are dichotomous or ordered-polytomous. We have represented all predictions by numerical scores, e.g., using the ordinal scale -1 = predicted non-carcinogen, 0.25 = predicted possible carcinogen, 0.5 = predicted probable carcinogen, 1 = predicted carcinogen, and ??? = no prediction (e.g., because data required to make a prediction were missing.)

In Figure 1, the single system that best predicts rodent carcinogenicity of the 44 test chemicals is that of Tennant, as also noted by Benigni (1995). If this system classifies a chemical as a rodent carcinogen, then there is a 71.4% chance (in this sample of 44 chemicals) that the long-term cancer bioassays will indeed show it to be a rodent carcinogen. Conversely, a chemical classified as a non-carcinogen has an 81.2% chance of not being a clear rodent carcinogen when the long-term bioassay is conducted. Thus, the total "resubstitution" error rate for the Tennant predictions alone is 25% = (28/44)(28.6%) + (16/44)(18.8%). Since these predictions are based on structural alerts learned from a training set of chemicals that did not include the 44 test chemicals considered in Figure 1, these estimates of performance are probably realistic. Resubstitution misclassification error rates for other individual predictive systems are as follows:

Tennant = 25%	Benigni = 34%	COMPACT = 41%
Weisburger = 32%	Bakale = 39%	DEREK = 41%
Tript 1 = 34%	CASE = 41%	TOPKAT = 43%
RASH = 34%	Tript 2 = 41%	Lijinsky = 45.5%

Table 1: Predictions and Results of Rodent Carcinogenicity Tests

NUMCHEM	MR	FR	MM	FM	CLASS	TENNANT	TRIPT1	BENIGNI	WEISBURGER	BAKALE	TOP_KAT	TRIPT	DEREK	COMPACT	LIJINSKY	MLT_CASE	DEREK_HY	RASH
1	NE	NE	NE	NE	NEG	-1	-1	-1	-1			1	-1	-1	-1	-1	-1	-1
2	NT	NT	NE	SE	POS	-1	-1	-1	-1	-1		-1	-1	1	-1	-1	-1	
3	EE	NE	NE	NE	EQV	-1	-1		-1	-1		-1			-1	1		0
4	NE	NE	NE	NE	NEG	-1	-1	-1	-1	-1		-1	-1	1	1	-1	-1	-1
5	NE	NE	NE	NE	NEG	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-0
6	NE	NE	EE	NE	EQV	-1	1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1
7	NE	NE	EE	EE	EQV	-1	-1		-1			-1	-1		-1		-1	1
8	NE	NE	NE	NE	NEG	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	1	-1
9	NT	NT	NE	NE	NEG	-1	-1	1	-1	1	-1	-1	-1	1	1	-1	1	-1
10	NE	NE	NE	NE	NEG	-1	1	-1	-1	1		1	-1	1	1	1	-1	-1
11	NE	EE	SE	NE	POS	1	1	0	-1	1	-1	-1	-1	-1	-1	-1	-1	1
12	CE	CE	CE	CE	POS	1	1	0	0	1	-1	1	-1	1	-1	1	-1	
13	SE	NE	EE	SE	POS	1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-0
14	SE	NE	NE	SE	POS	1	1	-1	-1	1	-1	-1	-1	1	1	-1	-1	0
15	NT	NT	NE	NE	NEG	1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1
16	SE	EE	EE	NE	POS	1	1		-1	-1		1			-1			0
17	NE	NE	SE	SE	POS	1	-1	-1	-1		-1	-1	1	-1	-1	-1	1	1
18	NE	NE	NE	NE	NEG	1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	1
19	NE	NE	NE	NE	NEG	1	1	-1	-1	-1	-1	1	-1	0	-1	1	-1	-1
20	EE	NE	SE	SE	POS	1	1	-1	-1		-1	1	-1	1	1	1	1	-1
21	EE	NE	NE	CE	POS	1	1	0	0	-1		-1	-1	-1	1	-1	-1	1
22	SE	EE	SE	NE	POS	1	1	0	-1	1	-1	1	-1	1	1	1	1	-1
23	NE	EE	NE	NE	EQV	-1	-1	0	-1			1	-1	1	-1		-1	
24	NE	NE	NE	NE	NEG	-1	-1	0	-1		1	-1	-1	1	-1	1	1	
25	NE	NE	NE	NE	NEG	-1	1	0	0	1		1	1	-1	1	1	1	
26	NT	NT	NE	NE	POS	-1	1	1	-1	1	-1	1	-1	1	-1	-1	1	-1
27	NE	NE	NT	NT	NEG	-1	1		-1			1	-1	-1	-1		1	-1
28	CE	CE	EE	EE	POS	-1	1	0	1	1	1	1	1	-1	1	1	1	-1
29	SE	NE	CE	CE	POS	1	1	0	-1			1	1	1	-1	1	1	-1
30	SE	SE	SE	NE	POS	1	1	1	-1	1	1	1	1	1	1	1	1	
31	EE	NE	NE	NE	EQV	1	1	1	-1		1	1	1	1	1	1	1	
32	NE	NE	SE	NE	POS	1	1	0	-1	-1	1	1	-1	1	1	1	-1	1
33	NE	EE	NE	NE	EQV	1	1	0	-1	-1	-1	1	-1	1	-1	-1	1	-0
34	SE	SE	CE	CE	POS	1	-1	1			1	-1	1	1	1	-1	1	
35	EE	EE	NT	NT	EQV	1	1			1		1	-1		-1		-1	-1
36	CE	CE	NT	NT	POS	1	1	1				1	1	1	-1	-1	1	
37	CE	CE	NT	NT	POS	1	1	0			1	1		1	-1	-1		
38	SE	EE	SE	CE	POS	1	1	0		1		1	-1	1	-1	-1	-1	1
39	CE	CE	CE	CE	POS	1	1	1	1	1		1	1	-1	1	1	1	1
40	CE	CE	NT	NT	POS	1	1	0	1	-1	1	1	-1	1	1	1	1	1
41	EE	NE	NE	NE	EQV	1	1	0	-1		-1	1	1	1	1	1	1	
42	NT	NT	EE	NE	EQV	1	1	1	-1	1	1	1	-1	1	1	1	1	0
43	CE	CE	CE	SE	POS	1	1	1	1	1	-1	1	-1	1	-1	1	1	0
44	CE	CE	CE	CE	POS	1	1	1	1	1	1	1	1	-1	1	1	1	0

Source: Bristol 1996

Clearly, no single test has a very low classification error rate, although 25% for the Tennant system is significantly lower than the error rate for other systems.

The classification tree in Figure 1 combines predictions from five different systems. For example, if both Tennant *and* Weisburger predict that a chemical is a rodent carcinogen, then so does the tree that hybridizes them (since 6 out of 6 cases predicted to be carcinogenic by Tennant and classified as carcinogenic or probably carcinogenic by Weisburger were in fact observed to be carcinogenic in the rodent bioassays). When Tennant predicts that a chemical is a rodent carcinogen and Weisburger does not, however, there is about a 64% chance that the chemical is carcinogenic. Introducing results of other tests, such as Bakala, Tript 2, and Lijinsky can help to further resolve this disagreement. Although not shown in Figure 1, among 14 chemicals classified as non-carcinogenic by both Tennant and Weisburger, two proved to be carcinogenic in the rodent bioassays.

Overall, hybridizing multiple predictions via the tree in Figure 1 leads to a slightly smaller average error rate (22.7% resubstitution error) than the best individual (Tennant) system alone has. Although tree model cross-validation with generalized degrees of freedom (Ye 1998), could yield better estimates of the true error rate for the tree by better correcting for bias due to over-fitting, the simple resubstitution error estimates suffice to illustrate the following key points:

e *Nonmonotonicity:* The best combination of expert system predictions need not be monotonic, i.e., the probability that a chemical is a rodent carcinogen may be decreased by learning that a system classifies it as such. For example, the Lijinsky node at the bottom of the tree shows a significant inverse relation between predicted and true classes, with the conditional probability that a chemical is a rodent carcinogen falling from 63.6% to 50%, given the results of the preceding tests in the tree, when the Lijinsky system predicts that it is a carcinogen. Although the sample sizes are small, the same phenomenon can occur in trees with more cases. The reason is that prediction errors made by different tests can interact in strong, potentially counter-intuitive ways. Exploiting such interactions enables classification tree hybrids to make improved predictions. This is a novel way to combine predictions from different sources, however, and it can violate many of the principles (such as unanimity and monotonicity) often proposed as normative axioms in previous approaches to combining predictions from different expert sources (Clemen 1989.)

High-order interactions among tests. The best predictions that can be achieved from a set of tests such as those in Table 1 may depend on interactions of many individual tests (e.g., five in Figure 1). Thus, the information contained in the one or two individually "best" tests does not subsume the useful information in the other tests.

o *Information value synergies:* The information value of a particular test can depend strongly on what other tests have been performed. For example, even though the predictive value of the Lijinsky test is only slightly better than random guessing when it is considered in isolation (i.e., its misclassification rate is close to 50%), it can help to identify high-probability carcinogens when used in the context of other tests in Figure 1.

F *No dominance:* It may be natural to think of some predictive systems as being strictly "better than" or "more informative than" others, so that when predictions from the best systems are known, those from less good systems should be ignored. Figure 1 suggests that relations among predictions can be more complicated, with even relatively weak predictive systems being able to add value for some combinations of predictions by the better systems. A formal basis for comparing predictive systems using classification trees is introduced in the next section.

That a combination of predictions from diverse sources can out-perform any of the individual sources is perhaps to be expected, based on much previous management science research on optimally combining or aggregating expert predictions (Clemen 1989). However, a novel feature of our approach is the use of classification trees, rather than analytic aggregation or averaging formulas, to combine the predictions from different AI and statistical prediction systems. This allows higher-order interactions among the prediction errors from different systems to be exploited in constructing combined predictions. As suggested by Figure 1, such interactions are potentially valuable: common combination methods based only on the variances and covariances of predictions from different sources may leave potentially valuable information unused. The tree approach also brings within a natural probabilistic framework systems that do not by themselves yield probabilistic predictions. This is done by treating their deterministic predictions (typically "carcinogen", "non-carcinogen" or "unable to make a determination") as values on which the combined, probabilistic prediction is conditioned.

3 METHODS AND DATA

Over the past three decades, a variety of increasingly sophisticated artificial intelligence and statistics methods have been brought to bear on the problem of predicting chemical carcinogens. Several well-developed approaches were recently evaluated in blind tests (Benigni 1996, Bristol 1996), i.e., they were used to predict whether various chemicals would be found to be rodent carcinogens in ongoing (typically, two-year) bioassay experiments in mice and rats. The predictions were published before the results of the experiments were known.

Key approaches tested for predictive accuracy include the following:

Structure-activity relation (SAR) programs. These consider physical and electronic properties, three-dimensional molecular structure, and molecular topological indices, indicating key invariants such as graph-theoretic structures (Perrotta 96) associated with DNA reactivity. The SAR and quantitative SAR (QSAR) approaches taken to date include:

q Benigni's method combines the structural alerts of Tennant and Ashby (below) with Bakale's coefficient of electrophilic reactivity, denoted Ke, (below) to obtain a QSAR score.

o CASE / MULTICASE (Cunningham 1998) is a Bayesian QSAR statistical expert system that uses statistically selected relations among attributes of chemical structures to identify substructures useful for predicting probable carcinogenicity. It differs from earlier QSAR expert systems in that it fully automates the selection of chemical substructures to be considered, rather than requiring a human user to select them from a library.

s COMPACT (Lewis 1998) is a QSAR system that calculates approximate molecular dimensions and molecular and electronic structures via "Computer-optimized molecular parametric analysis for chemical toxicity" to predict whether a chemical will be metabolically activated to a carcinogenic chemical by specific enzymes.

s DEREK (Marchant 1996) is a rule-based expert SAR system based on "deductive estimation of risk from existing knowledge" obtained from expert chemists.

n TOPKAT (Enslein 1990) applies statistical regression and discriminant analysis to chemical structural attributes to obtain SAR rules.

a Bakale Ke (Bakale 1992). This uses a single measured parameter, Ke = chemical electrophilic reactivity, to predict carcinogenic potential.

p Weisburger (an unpublished SAR system encoding expert intuition)

Activity-activity relation (AAR) programs. These use the spectrum of biological responses in relatively inexpensive assays (e.g., bacteria mutation tests or short-term toxicity and tests) to predict biological activities in more expensive and relevant systems (e.g., two-year rodent cancer bioassays). Some AAR systems, including that of Tennant and Ashby, also use any available data from previous cancer bioassays.

AAR systems evaluated for predictive validity (Benigni 1996, Bristol 1996) include:

) Tennant and Ashby's AAR system (Tennant 1990). This uses correlations among attributes of chemical structures, short-term mutagenicity test results (e.g., in *Salmonella*), rodent subchronic toxicity outcomes, and carcinogenicity test results if available. It identifies "structural alerts" indicating possible carcinogenicity. This system has been fine-tuned by its expert authors based on reviews of biological response profiles and chemical structures for over 300 chemicals. It incorporates much of their intuition. The system requires a human expert, i.e., it is not fully automatic.

i RASH (Jones 1996). The "rapid screening of hazards" method predicts carcinogenic potential based on the observed relative potencies of tested chemicals in different short-term bioassays. It is not fully automatic, but instead requires a human expert to select relevant comparisons.

t TRIPT (Bahler 1993) performs "tree and rule induction for predictive toxicology" via the machine learning algorithm C4.5, applied to the factors considered in the Tennant system.

c PROGOL (King 1996) applies inductive logic programming (ILP) to relational descriptions of chemical structures to induce simple, interpretable rules for SAR structural alerts.

Other methods for which predictions have been recorded include Fuzzy Adaptive Least Squares (Moriguchi 1996) and the unpublished predictive systems of Lijinsky and Weisburger.

Table 1 summarizes data on the outcomes of the different prediction methods (Bristol 1996) so that other AI and statistics researchers can try their own programs on the chemical carcinogenicity prediction task. It consists of the predictions for 44 chemicals made by different prediction systems using the above methods, ranging from statistical (e.g., TOPKAT) to rule-based expert systems to machine-learning (e.g., TRIPT)

approaches. Table 1 also presents the actual carcinogenicity outcomes observed for each chemical in both sexes of both mice and rats, based on experiments completed after the predictions were made. The codes for bioassay outcomes in individual species (M = mouse, R = rat) and sexes (M = male, F = female) are: CE = clear evidence of carcinogenicity; SE = some evidence; EE = equivocal evidence; NE = negative evidence

We analyzed these data, using the main classification tree algorithms implemented in KnowledgeSeeker™ with automatic Bonferroni adjustments to protect against multiple testing bias (Biggs 1991), to construct several trees that yield improved predictions. First, at the risk of over-training on the sample data, we conducted exploratory analyses of the whole data set (using KnowledgeSeeker's "Exhaustive" tree-growing algorithm) to detect patterns in errors across the different prediction methods. Then we used various random partitions of the 44 chemicals into training and test sets to assess the performance of the tree hybridization approach. (Typically, we used 29 chemicals to train and 15 to test.) A best-informed (S*) tree, as defined in the following section, was used to generate "hybridized" predictions from predictions already made by the different methods, along with statistics on their errors in the test set.

4 RESULTS AND DISCUSSION

Two major practical goals of new efforts in this field are to reduce the costs and increase the accuracy of predictive classification of chemicals. The costs are driven largely by *in vivo* testing implying that SAR and QSAR methods tend to be much less expensive than AAR methods, especially when the latter involve results of lengthy *in vivo* toxicity tests. By assembling batteries of tests that place relatively inexpensive tests first, the expected costs of reaching a classification decision with a specified level of confidence can sometimes be dramatically reduced. Indeed, this principle has been used in recent algorithms and heuristics for minimizing average costs of testing (Cox 1994). On the other hand, incorporating a few very expensive tests, such as a long-term cancer bioassay for a single rodent species and sex, can lead to dramatic improvements in accuracy if the predictive tests are used to select the test species and sex and to help interpret the results. The following paragraphs present our main findings on how classification trees can be used to improve cost-accuracy trade-offs and to compare different prediction systems.

4.1 REDUCING CLASSIFICATION COSTS

Figure 2 shows a tree with a resubstitution error rate of only 4.5%, far less than the 22.5% achievable if no long-term animal cancer bioassays are used (see Figure 1). It illustrates the value of combining the predictions from several systems with the results of a single long-term bioassay (either MM = male mice or FR = female rats), where the Tennant system's prediction is used to select which animal bioassay to perform. After performing the tests indicated in this tree, additional bioassays in other sexes or species do not improve predictive accuracy further.

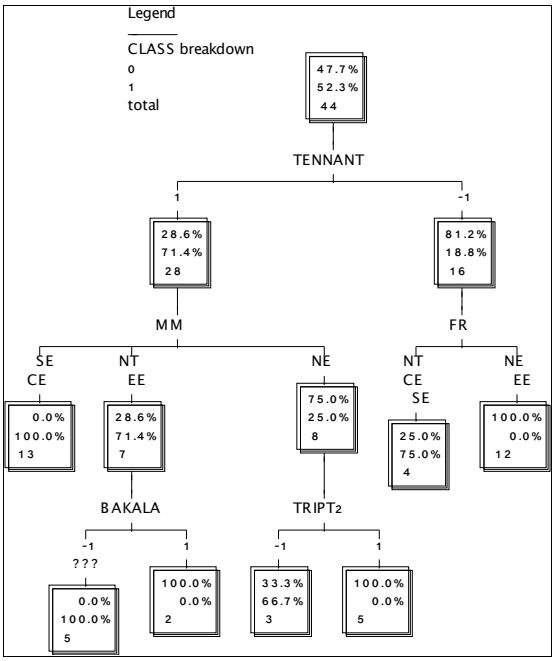

Figure 2: Predictions Help to Economize on Bioassays

Thus, classification tree analysis reveals that a chemical's carcinogenicity class is *conditionally independent* of the remaining information (predictions from other systems and bioassay experiment outcomes), at least as far as the tree-growing algorithm can discover, once one of the seven leaf nodes in Figure 2 has been reached.

It is noteworthy that the resubstitution error rate from the male mouse (MM) bioassay alone is 20%, while from the female rate (FR) test alone it is 30%. From the two together, it is 14%. Yet, hybridizing these tests with the other imperfect predictions in Figure 2 (Tennant, Bakala, and Tript2), which taken together have a joint error rate of 25%, produces a hybrid classification scheme with an error rate of less than 5%. Interpretively, this suggests that the MM and FR tests provide information that is approximately orthogonal to (complementary to) the information provided by the set of tests {Tennant, Bakala, Tript2}. This interpretation is strengthened by observing that, after constructing a tree using these three variables, the MM and FR tests will still enter at the bottom of that tree if allowed to, implying that a chemical's class is *not* conditionally independent of MM and FR, given the predictions from

{Tennant, Bakala, Tript2}. Once MM and FR have entered, however, no other variables in Table 1will, showing that the set {Tennant, Bakala, Tript2, MM, FR} is *sufficient* for the full set of variables. Searching the set of possible trees shows that this is a *minimal sufficient set* (i.e., none of its subsets has the sufficiency property) and that no other set achieves both a smaller error rate and a smaller cost (assuming that animal bioassays cost more than AAR methods and that AAR methods cost more than SAR or QSAR methods.) See (Cox 1994) for search algorithms and heuristics for finding cost-effective trees.

4.2 COMPARING PREDICTIVE SYSTEMS

The classification tree framework for combining predictions contributes a new technique for comparing information sources. Roughly speaking, one source of predictions may be considered "better-informed" or "more valuable" than another if *every* rational decision-maker would prefer to obtain an observation from the first instead of the second before making a decision. (This assumes that the payoff or utility from the decision depends on the true state, e.g., carcinogenic or not, for the chemical being classified.) This comparative binary relation generally yields a partial ordering of information sources when the characteristics of the sources (i.e., probabilities of outputs given the true states) are known. While several equivalent characterizations have been given for determining when one information source is more valuable than another, the classification tree framework provides a simple test that does not require a priori knowledge of the probability characteristics of the sources. Call source S1 *at least as well informed as* source S2 if the classification of a chemical is conditionally independent (CI) of the prediction from S2, given the prediction from S1. S1 is *better informed* than S2 if S1 is at least as well informed as S2 but not *vice versa.* The classification tree test for this relation is as follows.

Comparing Predictive Systems via a Tree Algorithm	
C	Take the correct classification of a chemical (e.g., carcinogen or not) as the dependent variable.
c	Split the correct classification on the output (predictions) from S1 to form a one-split tree. Call this tree T1.
t	Allow tree T1 to be extended by splitting each of its leaves on the predictions from S2.
l	If S2 does not enter the tree when this is done, but if splitting the chemical classification first on S2 and then on S1 results in S1 entering the tree below S2, then S1 is better informed than S2. Interpretively, S1 is "sufficient for" S2, but not *vice versa;* see DeGroot, 1970, Chapter 14.

This characterization of comparative expertise defines a partial ordering that complements earlier ones in the statistical decision theory literature (DeGroot, 1970, 433-439). It can readily be extended to compare *subsets* of variables (or tests, or predictive systems), by treating each subset in turn as the allowed set of independent variables entering a classification tree analysis. For example, after conditioning (by growing a tree, T1) on {Weisburger, RASH, COMPACT}, Tennant and Tript2 will still enter at the bottom of the tree if allowed to. Conversely, after conditioning on {Tennant, Bakala, Tript2}, Weisburger, RASH, and COMPACT will all still enter if allowed to. Thus, neither set is more informative than the other: they are complementary.

4.3 OPTIMALLY COMBINING PREDICTIONS

The ability to compare prediction systems based on trees suggests constructing a new, *best-informed source of* predictions, say, S*, by combining the predictions from individual sources into a tree such no other tree composed from these sources is better-informed than S*. A useful approximate construction heuristic that uses standard classification tree algorithms is the following. Grow an initial tree myopically, e.g., by choosing the strongest predictors of correct classes in the training set first. Then refine it by making pairwise swaps of variables below the root node with the root node variable until no further improvements (defined as reduction of average prediction error rate in the test set) can be found. Apply this tree-improvement routine to several initial trees formed by random selection of the top few candidate splits at each node. The result is often a tree that (a) yields the smallest achievable average prediction error in the test set; and (b) does so using an efficient set of variables, i.e., a set of variables such that any proper subset yields significantly deteriorated performance. This tree is a practical estimate or approximation to S*, the best-informed predictor of chemical class that can be formed from the sources considered.

In our experience, as discussed in Section 2, the best-informed source if often inconsistent with axioms that have sometimes been proposed in the management science literature for combining expert predictions (e.g., the "unanimity" axiom, according to which S* should make the same prediction as its component sources when all of them agree). However, it is easy to demonstrate by examples that, in these cases, the axioms are not useful, whereas S* is, in making the best possible predictions. Thus, it appears that classification trees may offer a useful general alternative to previous methods of combining expert predictions, as well as making more specific contributions to chemical carcinogen prediction.

4.4 OTHER FINDINGS

Other findings from our classification tree analysis of Table 1 include the following.

1 The different methods are much better predictors of each other than of the true classification of the chemicals. This suggests that there is important information captured in the rodent cancer bioassays that is not captured in the predictive methods currently in use. As shown in Figure 2, such complementary information can be exploited to reduce the number of expensive bioassays performed.

e Combining predictions from different predictive methods leads to a best-informed tree that improves on the predictive accuracy of any single method. The best-informed tree is not unique, however.

b When only the least expensive (SAR or QSAR, but not AAR) tests are considered, the best hybrid tree classifier based on {Bakala's Ke, MULTI-CASE, COMPACT, DEREK, TOPKAT} has a 20% resubstitution misclassification error rate.

r A key goal of carcinogen prediction has been to identify a battery of low-cost tests and assays that would collectively be as informative as the much more expensive rodent bioassays about the likely carcinogenicity of chemicals. Classification trees were applied to test how well this goal has been achieved. Whether carcinogenicity in each species and sex is conditionally independent of carcinogenicity in the other three, given the results of all predictive methods, can be tested by the two-phase tree-growing procedure outlined above. It turns out that the other carcinogenicity bioassays contain relevant information not captured in the predictive methods (i.e.., the predictive methods being used are not sufficient for the carcinogenicity experiments.)

n On the positive side, classification trees show that carcinogenicity of a chemical in a specific rodent species and sex can be predicted as well from carcinogenicity testing results in one other species and sex and a few (typically two) of the prediction methods as it can be from results of carcinogenicity testing in all three other species-sex combinations

t The classification tree method suggests the possibility (and provides a constructive algorithm) for combining whole-animal carcinogenicity testing with less expensive predictive methods to obtain predictions of human carcinogenicity that are at least as informed as methods based on more extensive whole-animal testing in additional species and sexes.

5 SUMMARY AND CONCLUSIONS

In summary, we have identified a tree-based approach to combining the results of multiple tests to reduce test costs (e.g., by using results of less expensive tests to determine which expensive ones to perform) and to reduce error rates by hybridizing predictions based on complementary information. The approach appears promising for the data in Table 1. It can be implemented using standard classification tree software. However, some important open issues remain. These include the following.

1. *Tracking concept drift*. The classification tree analysis revealed that year of completion of peer review of rodent cancer bioassays is itself quite informative about the likelihood that a chemical is a rodent carcinogen. Roughly 30% of chemicals reviewed in 1990, 60% of those reviewed in 1991-1993, and 100% of the (three) chemicals reviewed in 1994 and 1995 were found to be rodent carcinogens. Thus, the proportion of rodent carcinogens among chemicals selected for long-term cancer bioassays may be increasing over time. (Indeed, if year of review is used as the sole predictor of rodent carcinogenicity, the sample misclassification error rate is 36%, lower than for several of the predictive systems.) When training and test sets are obtained by partitioning chemicals according to the year in which a peer-reviewed cancer bioassay was completed, it appears that the first chemicals tested (mainly genotoxic ones) yield trees that are especially weak predictors of the carcinogenicity of later chemicals (which contain more non-genotoxic ones). Thus, when the concept being learned "drifts" over time (e.g., away from genotoxic and toward non-genotoxic carcinogens, in this case), it is important to make sure that the training set is balanced (or re-balanced) to adequately emphasize the components that are to be predicted.

2. *Formal cost-optimization.* This paper has emphasized construction of best-informed sources from several less-informed sources. As briefly mentioned, a useful extension would be to assign costs to the different tests and seek a *minimum-expected cost tree* that balances the costs of testing against the costs of decision error. Computational complexity results and practical heuristics are available for such problems (Cox 1994).

3. *Latent variables.* A potentially desirable approach to predicting chemical carcinogenicity is to allow for hierarchical concept-learning, including induction of latent variables (such as "genotoxic carcinogen"). Such variables do not arise as

Boolean combinations of attribute values, but may greatly simplify the interpretation of attribute value combinations. It may be worthwhile to extend classification tree algorithms to partition training sets into relevant and irrelevant exemplars, based on hypothesizing a latent variable (e.g., the "genotoxic" classification) that is related to the observed attribute values but not directly measured. For example, we have found that classification trees can provide powerful predictors of mineral oil carcinogenicity, with clear advantages compared to older statistical methods, if latent variables can first be used to partition the training and test sets into relatively homogeneous subsets.

Acknowledgment

This research was stimulated by and has benefited from many discussions with Dr. Michael Bird of Exxon Biomedical Science, Inc. (EBSI). I am grateful to Dr. Bird and to EBSI for encouraging and supporting my research on better ways to use information about biological response profiles to predict the likely health effects of chemicals.

References

Ashby, J., and D. Paton, 1993. The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogens. *Mutation Research,* 286, 3-74.

Bahler D. and D.W. Bristol, 1993. The induction of rules for predicting chemical carcinogenesis in rodents. *Ismb*; **1**:29-37

Bakale G. and R.D. McCreary R.D., 1992. Response of the ke test to NCI/NTP-screened chemicals. II. Genotoxic carcinogens and non-genotoxic non-carcinogens. *Carcinogenesis*; **13** (8): 1437-45

Benigni, R., 1997. The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: definitive results. *Mutation Research*; **387**(1):35-45

Benigni, R., 1996. Predicting chemical carcinogenesis in rodents: The state of the art in light of a compara-tive exercise. *Mutation Research,* 334, 103-113.

Benigni, R., 1991. QSAR prediction of rodent carcinogenicity for a set of chemicals currently bioassayed by the US National Toxicology Program.
Mutagenesis 1991 Sep;6(5):423-5

Biggs, D., B. de Ville, E. Suen, 1991. A method of choosing multiway partitions for classification and decision trees. *J. Applied Statistics,* 18, 1, 49-62.

Bristol D.W., J.T. Wachsman, A. Greenwell, 1996. The NIEHS Predictive-Toxicology Evaluation Project. *Environ. Health Perspectives;***104** Suppl 5:1001-10.

Chevalier, S., R.A. Roberts RA, 1998. Perturbation of rodent hepatocyte growth control by nongenotoxic hepato-carcinogens: mechanisms and lack of relevance for human health. *Oncology Reports*; 5(6):1319-27.

Clemen, R.T., 1989. Combining forecasts: A review and annotated bibliography," (with discussion). *International Journal of Forecasting,* **5**, 559-583.

Cox, L.A., Jr., 1997. Does diesel exhaust cause human lung cancer? *Risk Analysis*, **17**, 6, 807-829.

Cox, L.A., Jr., and Y. Qiu, 1994. "Minimizing the expected costs of classifying patterns by sequential costly inspections," in Cheeseman, P. and R.W. Olford (Eds), *Selecting Models from Data.* Springer-Verlag, *Lecture Notes in Statistics,* Volume 89, pp. 339-350. New York.

Cunningham AR, G. Klopman, H.S. Rosenkranz, 1998. Identification of structural features and associated mechanisms of action for carcinogens in rats. *Mutation Research* **31**; 405(1):9-27.

DeGroot, M.H., 1970. *Optimal Statistical Decisions.* McGraw-Hill.

Enslein K., B.W. Blake, H.H. Borgstedt, 1990. Prediction of probability of carcinogenicity for a set of ongoing NTP bioassays. *Mutagenesis*; **5**(4): 305-6

Gold, L.S., T.H. Slone, N.B. Manley, L. Bernstein, 1991. Target organs in chronic bioassays of 533 chemical carcinogens. *Environmental Health Perspectives*, **93**, 233-246.

Jones T.D. and C.E. Easterly, 1996. A RASH Analysis of National Toxicity Program Data: Predictions for 30 Compounds to Be Tested in Rodent Carcinogenesis Experiments. *Environmental Health Perspectives*; **104** S (5): 1017-30

King, R.D., and A. Srinivasan, 1996. Prediction of rodent carcinogenicity bioassays from molecular structure using inductive logic programming. *Environmental Health Perspectives,***104**S(5): 1031-1039.

Lewis D.F., C. Ioannides, D.V. Parke, 1998. Further evaluation of COMPACT, the molecular orbital approach for the prospective safety evaluation of chemicals. *Mutation Research*,**13**;412(1):41-54