

# Learning to Predict Carcinogenesis of Unstudied Chemicals in Rodents from Completed Rodent Trials

**Carol A. Wellington and Dennis R. Bahler**

Artificial Intelligence Laboratory  
Department of Computer Science  
North Carolina State University  
Raleigh, NC 27695-8206

## Abstract

The National Toxicology Program (NTP) studies chemicals to determine if they are carcinogenic. These experiments include subchronic (90 day) and chronic (2 year) rodent exposures studies and, therefore, are costly and time consuming. The long-range goal of our research is to learn Bayesian belief networks from the NTP data and use them to predict the classification of chemicals at various milestones during the process when that information could be used to justify either continuing or terminating the experiments.

NTP has data from 226 chemicals which have previously been classified. The data contain almost 1000 attributes, including the results of microbial assays, physical-chemical parameters, and the results of 90-day rodent exposure studies. While the data set contains continuous, discrete, and binary attributes, the majority of the attributes (836) are the subchronic exposure study results representing the presence or absence of organ pathology. Detection of a particular combination of organ and morphology (damage) is rare, so these attributes are very sparsely positive. This makes detecting significance of an attribute difficult. In addition, not all of the exposure studies are done for every chemical, so a number of chemicals are missing the attribute values for many of these attributes.

Our approach to handling this complex data set has been to use various feature selection techniques and statistical analysis (such as linear discriminant analysis) on the exposure study result attributes. The models we build use the results of that analysis in combinations with the other attributes to predict various subsets of the chemical population. The cross-validated accuracies of these models have ranged from 70% to 92%. We are also building models from the data set directly.

In analyzing the results of our preliminary models, it became apparent that much of the difficulty with this data set comes from the level of noise of the result attributes in the exposure studies. Our endpoint is a combination of many different types of carcinogenicity where each is likely to show different types of organ damage, so our model is called upon to find many biological pathways to many endpoints. It is our opinion that this wide endpoint is aggravating the noise in the data. In future experiments, we plan to replace this endpoint with endpoints for cancers in specific organs, the goal being to increase the accuracy and intuitiveness of our models.

# 1 Introduction

The National Toxicology Program (NTP) evaluates test agents (generally chemical compounds or mixtures) to determine their carcinogenicity. These evaluations involve various *in vitro* and *in vivo* experiments whose results are evaluated to determine the level of carcinogenicity. That result is called the **overall bioassay result**. The results of the *in vitro* experiments include information such as the mutagenicity of the test agent to salmonella. The *in vivo* experiments include short-term and long-term rodent exposure studies. The short-term studies are 90 days long and are used to determine dose levels for the long-term studies as well as short term or **subchronic** effects of the test agent. The long-term studies typically last 2 years and determine longer term or **chronic** effects of the test agent [5]. These *in vivo* experiments are expensive, so alternatives to these techniques are being actively sought by many investigators including us.

This paper details two models which are used to predict the overall bioassay result of an NTP carcinogenicity study. The techniques glean information from data gathered about previous NTP carcinogenicity studies and use that information to predict the results of incomplete NTP studies. Since the goal is to provide an alternative to the long term rodent exposure experiments, only information from the *in vitro* and subchronic *in vivo* studies are used.

The rest of this section of this paper describes the training set used in the two models and an overview of those models. The Methods section describes the construction of the models and justifies the decisions made in building those models. The Results section contains an evaluation of the two models.

## 1.1 The Training Set

The techniques described in this paper use information from previous NTP carcinogenicity studies, called the **training set**. Our training set consisted of data from 191 test agents bioassayed by NIH as reported in NTP Technical Reports 200 to 458. The NTP studies result in five classifications: clear evidence, some evidence, equivocal evidence, negative evidence and inadequate study. In our training set, clear evidence and some evidence results were combined into one class called positive because a number of the test agents in the training set were classified before NTP subdivided positive results into the two classes. Our training set did not include test agents which resulted in inadequate study classifications, as those studies had been deemed to be incomplete or flawed in some manner. Test agents whose overall bioassay result was classified by NTP as equivocal were also eliminated, as an

equivocal result is not really positive or negative, and therefore there is some debate about what that result means. The 191 remaining test agents in the training set had 128 positive overall bioassay results and 63 negative bioassay results.

For each test agent, the training set contains the values of attributes which are inherent to the test agent, which we call the **agent-specific attributes**, as well as the results of up to four 90-day rodent exposure studies which we call the **experiment-specific attributes**. The agent-specific attributes are pieces of information about the test agent which can be gathered prior to the 90-day exposure studies. These are summarized in Table 1.1 and include the results of salmonella mutagenicity studies, 20 physical chemical attributes, and presence of those chemical structural alerting submolecules labeled as a through u in [1].

Structural Alerts (a-u)	the number of times each structural alert is present
Salmonella Mutagenesis	32 possible values encoding the results of all salmonella studies
Clogp	octanol/water partition coefficient
Dipole	dipole moment
Ehomo	highest occupied molecular-orbital energy (a.u.), output from Gaussian single point calculation
Electronegativity	electronegativity value
Elumo	lowest occupied molecular-orbital energy (a.u.)
Logp	estimate of partition coefficient
Hardness	molecular hardness
Ovality	molecular ovality (unitless)
Volume	molecular volume
Weight	molecular weight value to the larger of 4 sig fig or 2 decimal places
PA	planar area
(PA/D)**2	square of the ratio of planar area to z-depth squared
Pka	dissociation constant, from ASTER
RA	rectangular area
(RA/D)**2	square of the ratio of rectangular area to z-depth
Rate Ke	Bakale's electron capture rate [2] experimentally determined or published values
Z depth	Z-depth or thickness

Table 1: Agent-Specific Attributes in the Training Set

The experiment-specific attributes contain the results of the 90-day rodent exposure studies. At the end of each 90-day study, the rodents are dissected and the presence of specific organ-morphology pairs is noted. The test agents in the training set contain evidence

of 72 morphologies in one or more of 38 different organs, so there are 2736 possible organ-morphology pairs. However, only 209 of these were found during the studies of test agents in the training set, so the training set contains only those 209 results for each experiment. The training set contains the results of up to four 90-day studies: one each for male rats, female rats, male mice, and female mice. Therefore, the training set contains  $4 \times 209 = 836$  experiment-specific attributes. These are referred to as the **subchronic attributes** and can be represented as a bit vector for each test agent where each bit represents the presence or absence of a specific organ-morphology result in a specific sex-species experiment.

While all of these agent- and experiment-specific attributes are present in the training set, there are many instances of missing values. In the agent-specific attributes, some of the physical chemical parameters are uncomputable or were not computed for some of the test agents. In the experiment-specific attributes, there are some test agents for which all four 90-day studies were not completed. In both of these cases, the values were left unspecified (missing) in the training set.

## 1.2 Overview of Two Classification Techniques

Classification is the task of dividing a set of objects into groups based on the values of one or more attributes of those objects. Those groups are called **classes**. For our training set, the test agents are divided into two classes by the overall bioassay result of the test agent: positive and negative. The goal of this research is to be able to predict the class of a test agent not in the training set from the values of the agent- and experiment-specific attributes for that test agent. We applied two techniques to this problem.

The first classification technique was linear discriminant analysis (LDA). LDA is a statistical technique which takes the values of a set of attributes and the classification for each test agent in the training set and finds a linear combination of the attributes for each possible classification. For each test agent, those linear combinations of the attributes give a **linear discriminant score** for each classification. A test agent's classification is predicted by the classification which has the highest linear discriminant score. LDA finds the linear combinations which will minimize the probability of misclassification within the training set.

LDA was chosen because it can handle sets of attributes like the subchronic and structural alert attributes. These sets of attributes are nominal attributes with very few attributes positive in any test agent and with each attribute positive in very few test agents. However, LDA has no method for handling attributes whose values are missing in the training set and, therefore, could not be applied to a number of the agent-specific attributes.

The second classification technique was learning Bayesian classifiers from the training set. A Bayesian classifier is a method for representing the joint probability distribution across the attributes and the classification. It is a graphical model which allows the joint probability distribution to be represented by conditional probabilities which can be learned from the training set. After it is built, a Bayesian classifier allows the probabilities that a test agent lies in each class to be computed from the values of the attributes for that test agent. There are techniques which allow Bayesian classifiers to be learned from data which contains missing values and from real-valued attributes, so we used Bayesian classifiers to combine the linear discriminant scores results from LDA with the agent-specific attributes in the training set.

Information from experts within the field of predictive toxicology suggests that the subchronic attributes may be the most predictive attributes. However, because there were many such attributes, the ratio of the number of subchronic attributes to the number of test agents in the training set was very high. This makes gleaning information from that portion of the training set quite difficult. Therefore, we used **Wilks' lambda** [7] to select a subset of the subchronic attributes which best discriminate the classes of test agents.

LDA was first applied to the subset of subchronic attributes which were selected by Wilks' lambda (we call these the **selected subchronic attributes**). While this works well for test agents which exhibited those subchronic results, it classified all test agents which do not exhibit those subchronic results as negative. To predict those test agents, we applied LDA to the structural alert attributes.

A Bayesian classifier was then built which combined the LDA scores for the subchronic attributes, the LDA scores for the structural alert attributes, the salmonella attribute, and the physical chemical attributes. As with LDA, this worked well for test agents which exhibited at least one of the selected subchronic attributes but predicted all test agents which did not exhibit any of the selected subchronic attributes as negative. Therefore, a second Bayesian classifier which did not include the LDA scores for the subchronic attributes was built to predict the classification for those test agents.

In summary, the two techniques of LDA and Bayesian classifiers were each applied to two subsets of the training set: test agents which exhibited at least one of the selected subchronic attributes and those which did not. The result was four separate models. When predicting the classification of a test agent not in the training set, the appropriate pair of models (one LDA and one Bayesian classifier) is used based on whether that test agent exhibited at least one of the selected subchronic attributes. The result is two separate predictions of the classification: one from the LDA model and one from the Bayesian classifier.

## 2 Methods

### 2.1 Feature Selection of Subchronic Attributes

The training set contains 836 subchronic attributes which are very sparsely positive. No test agent has more than 15 subchronic attributes positive and no subchronic attribute is positive in more than a very few test agents. The number of attributes and the sparsity of the positive values of these attributes make gleaning statistically significant information from the attributes difficult. Therefore, we used Wilks' lambda to choose a subset of subchronic attributes would predict the overall bioassay result with sufficient accuracy. Wilks' lambda represents the likelihood that the sample mean predicts the population mean and is defined as the ratio of the generalized variance of a subset of the attributes to the generalized variance of all of the attributes. As Wilks' lambda approaches 1, the distribution of the subset of attributes approaches the distribution of all of the attributes. Thus comparing the Wilks' lambdas after the addition of each remaining attribute measures the discriminatory power of the new model.

Beginning with no attributes, we attempted to add attributes to the model one at a time. At each step, the Wilks' lambda was calculated for the set of previously chosen attributes plus each of the remaining attributes. At each step in feature selection, the attribute which contributed most to the discriminatory power of the model was added to the model.

The addition of variables was iterated until no attribute contributed significantly to the discriminatory power of the model. The criterion for stopping the addition of attributes was a significance level of 0.75. This significance level seems quite high, but was necessitated by the extreme sparsity of the data. A high significance level criteria was necessary to select enough attributes to allow reasonable discriminatory power in the resulting model. This resulted in the selection of 86 subchronic attributes which are summarized in Table 2.

It is important to note that test agents which were missing the values for subchronic attributes were not included in this analysis since Wilks' lambda has no means for dealing with missing values.

### 2.2 Models Using Linear Discriminant Analysis Only

#### 2.2.1 Linear Discriminant Analysis

The first technique we used was Linear Discriminant Analysis (LDA). LDA is a statistical technique which finds linear combinations of a set of attributes which will best distinguish classes of examples. The linear discriminant score for an example belonging to class  $i$  is [7]:

$$d_i(x) = \mu_i' \Sigma^{-1} x - \frac{1}{2} \mu_i' \Sigma^{-1} \mu_i + \ln p_i$$

where:

- $\mu_i$  is a vector of the true means of each attribute within class  $i$ ;
- $\mu_i'$  is the transpose of vector  $\mu_i$ ;
- $\Sigma$  is the true covariance matrix;
- $x$  is the vector of attributes for the specific example;
- $p_i$  is the prior probability of class  $i$ .

This linear discriminant score can be estimated by:

$$\hat{d}_i(x) = \bar{x}_i' S_{pooled}^{-1} x - \frac{1}{2} \bar{x}_i' S_{pooled}^{-1} \bar{x}_i + \ln p_i \quad (1)$$

where:

- $S_{pooled}$  is the pooled covariance matrix;
- $\bar{x}_i$  is the average of attributes for examples in class  $i$ ;
- $\bar{x}_i'$  is the transpose of  $\bar{x}_i$ .

Using these definitions, LDA will allocate an example to the class which has the highest  $\hat{d}_i(x)$  for that example. Choosing this definition of  $\hat{d}_i(x)$  and this allocation rule can be shown to minimize the total probability of misclassification, assuming that the classes are multivariate normal distributions with equal covariance matrices.

Considering Equation 1 further, we can see that it reduces to a linear combination of the attributes of the example. LDA finds the coefficients for a linear combination for each class so the resulting scores will best distinguish between the classes.

Since we are going to have a different linear combination of the attributes of an example for each class, LDA can be viewed as multiplying the vector of the example's attributes by a matrix containing the coefficients for all of the classes. The result of that multiplication is a vector of linear discriminant score estimates (one for each class).

### 2.2.2 Model Using LDA on Subchronic Attributes

LDA was first applied to the subchronic attributes for each test agent in the training set. The result was two scores per test agent: the discriminant score for negative and the discriminant score for positive. The classification of a test agent was predicted by the higher of the two

scores. In calculating the discriminant scores for test agents which were missing the values of any of the subchronic attributes, they were considered to be present, but negative (not exhibited) for that test agent. Since a positive value for a subchronic attribute was presumed to be evidence of carcinogenicity, assuming that missing attribute values had not occurred was not likely to make us predict that a test agent will be carcinogenic. Since these attributes are rarely positive, this assumption seems reasonable.

The application of LDA to the subchronic attributes was cross-validated on test agents which were not missing the values of any of the selected subchronic attributes. The results of that analysis are summarized in Table 2. While its accuracy of 79% was promising, many of its errors resulted from the fact that this model predicted as negative all of the test agents which did not exhibit any of the selected subchronic attributes. Of the 155 test agents which were not missing the values of any of the selected subchronic attributes, 29 had none of those attributes positive. Of those 29 test agents, 17 were classified by NTP as negative and 12 as positive. Clearly, these attributes do not allow these test agents to be distinguished between the classes. Therefore, they were predicted using a separate LDA model. Removing them from the accuracy measurement of this model is shown in Table 3 and raises the accuracy of this model to 84%.

		Predicted Class		
		Negative	Positive	Total
Actual Class	Negative	49	4	53
	Positive	28	74	102
	Total	77	78	155

Table 2: Cross-Validated Results of DA For TestAgents Not Missing Subchronic Attributes

		Predicted Class		
		Negative	Positive	Total
Actual Class	Negative	32	4	36
	Positive	16	74	90
	Total	48	78	126

Table 3: Cross-Validated Results of LDA For Test Agents With Selected Subchronic

### 2.2.3 Model Using LDA on Structural Alert Attributes

For test agents which did not exhibit any of the selected subchronic attributes, a separate application of LDA was made using the structural alert attributes. The results of cross-validating that model are shown in Table 4. Since the accuracy of this model was only 58%, it represents a minimal improvement in accuracy for test agents which exhibited none of the selected subchronic attributes (predicting them using LDA on subchronic attributes had an accuracy of 57%), and should only be used for those test agents.

		Predicted Class		
		Negative	Positive	Total
Actual Class	Negative	21	4	25
	Positive	16	7	23
	Total	37	10	48

Table 4: Cross-Validated Results of LDA on Structural Alert Attributes (all test agents)

## 2.3 Models Using Bayesian Classifiers

While LDA achieved reasonable accuracy using subchronic and structural alert information, it could not deal well with attribute values which were missing from many test agents. Therefore, the technique of Bayesian classifiers was used to combine the results of LDA on the subchronic and structural alert attributes with the other agent-specific attributes.

### 2.3.1 Bayesian Classifiers

We chose to start by constructing only naive Bayesian classifiers. A naive Bayesian classifier is a Bayesian classifier with one node for the class attribute and one node for each of the other attributes. There is an edge from the class node to each of the attribute nodes and there are no other nodes in the graph. Therefore, once the features have been selected, they define the structure of the belief network.

## 2.4 Learning the Probabilities in the Bayesian Classifier

Because there are probabilities at every node and conditional probabilities at every arc, there are many probabilities which must be specified in building a network. However, once the

structure of the network is specified, all of these probabilities can be learned from a set of examples from the population being modeled.

For our experiments, the Bayesian belief network application Netica was used to learn the probabilities. Netica assumes the conditional probabilities being learned are independent and that the prior distribution is Dirichlet. It then uses a beta function, parameterized by experience and a probability number, to represent the distribution over possible probabilities.

Netica keeps one experience node representing the number of examples seen for each possible parent configuration at each node. With that experience, it maintains one probability for each state of the node. Initially, all probabilities are equal. As each training example is seen, nodes which have values specified for themselves and their parents (which in this case is the class which is never a missing value) have their experience and probability vectors updated as follows:

$$E'_s = E_s + 1$$

$$P'_s = (P_s * E_s + 1) / E'_s$$

$$P'_{\bar{s}} = (P_{\bar{s}} * E_s) / E'_s$$

where:

$E_s$  is the experience for the state of the parents in the current case;

$E'_s$  is the updated experience for the state of the parents in the current case;

$P_s$  is the probability the node will be in the state given in the current case prior to learning this case;

$P'_s$  is the updated probability the node will be in the state given in the current case;

$P_{\bar{s}}$  is the vector of probabilities for states other than the state given in the current case;

$P'_{\bar{s}}$  is the updated vector of probabilities for states other than the state given in the current case.

Note that the probabilities for states not in the current case are updated to keep them normalized.

### 2.4.1 Discretizing Continuous Attributes

Each node of a Bayesian classifier must have a finite number of states, so a state of a node representing a continuous valued attribute cannot be associated with a particular value of that attribute. For nodes representing continuous attributes, each state of the node is associated with a subrange of the possible values. Finding those subranges is called

**discretizing** the attribute because it allows a finite number of states to be associated with selected, discrete subranges of the possible values.

Discretization was accomplished using an algorithm based on a minimal entropy heuristic [3, 4]. The entropy of a set S of instances (in this case, test agents) is

$$Ent(S) = - \sum_{i=1}^k P(C_i, S) \log(P(C_i, S))$$

which roughly measures the amount of information required to specify the classes in S. The more heterogeneous is S with respect to class, the larger its entropy will be. The goal of the algorithm is to take a set of instances S (which in this case is the test agents in the training set) and partition it into subsets based on ranges of the attribute being discretized so that the entropy of the subsets is minimized.

The algorithm takes a set of instances (in this case test agents) and sorts them by the attribute being discretized. The algorithm then looks for at each possible partition boundary T which will divide S into two subsets  $S_1$  and  $S_2$  and measures the **class information entropy of the partition induced by T**,  $E(T, S)$  which is defined as:

$$E(T, S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

The algorithm will discretize the attribute at that partition boundary  $T_{min}$  which minimizes  $E(T, S)$  and will then recursively discretize the subsets  $S_1$  and  $S_2$ . This recursion will stop when partitioning a subset of the test agents does not result in sufficient entropy gain. Entropy gain of a partition boundary T is measured by

$$Gain(T, S) = Ent(S) - E(T, S)$$

and recursive partitioning will stop when the value for  $Gain(T_{min}, S)$  passes below a predetermined threshold.

#### 2.4.2 Bayesian Classifier For Test Agents With Selected Subchronic Results

The first Bayesian classifier we built was designed to predict the class for test agents which exhibited at least one of the selected subchronic attributes. The evidence nodes for this network included nodes for all of the agent-specific attributes except the structural alert attributes, the structural alert LDA scores and the subchronic LDA scores. Interestingly, the automatic discretization algorithm did not partition any of the physical chemical attributes or the structural alert LDA score for negative (in other words, all values were put into one

state of the node), so these attributes were not used by the network and were subsequently removed. In addition, the presence of the salmonella attribute lowered the cross-validated accuracy of the model, so it was removed. The resulting network included evidence nodes for only the subchronic LDA scores and the structural alert LDA score for positive.

The results of cross-validating this model on the test agents which exhibited at least one of the selected subchronic attributes is shown in Table 5. On that set of test agents, this network achieved a cross-validated accuracy of 92%.

		Predicted Class		
		Negative	Positive	Total
Actual Class	Negative	37	4	41
	Positive	8	95	103
	Total	45	99	144

Table 5: Cross-Validated Results of Bayesian Classifiers For Test Agents With Selected Subchronic Attributes

### 2.4.3 Bayesian Classifier For Test Agents With No Selected Subchronic Attributes

As with the LDA models, our first Bayesian classifier predicted all test agents which exhibited none of the selected subchronic attributes as negative, so we built a second Bayesian classifier which used the structural alert LDA scores and all of the agent-specific attributes except the structural alert attributes. As with the first Bayesian classifier, none of the physical chemical attributes and the structural alert score for negative were partitioned during discretization, so they were removed from the network. In addition, the presence of the salmonella attribute lowered the cross-validated accuracy of the model, so it was removed. This left only one evidence node in the network: the structural alert LDA score for positive.

The results of cross-validating this model on the test agents which exhibited none of the selected subchronic attributes are shown in Table 6. This model achieved a cross-validated accuracy of only 72%, so it was used only for test agents in this group.

		Predicted Class		
		Negative	Positive	Total
Actual Class	Negative	23	0	23
	Positive	13	11	24
	Total	36	11	47

Table 6: Cross-Validated Results of Bayesian Classifier For Test Agents Without Selected Subchronic Attributes

### 3 Results

#### 3.1 Predictive Value of LDA Models

There are a number of statistics which are used to measure the ability of a model to predict a classification. **Positive predictivity** is the percent of the predictions the model made to the positive class which were correct and **negative predictivity** is the percent of the predictions the model made to the negative class which were correct. These give a measure of how likely it is that a prediction of a specific class is correct. **Sensitivity** is the percent of positive examples which were correctly classified and **specificity** is the percent of negative examples which were correctly classified. These give measures of how well the model can predict test agents of each class.

Table 7 combines the results of the two LDA models. Table 8 summarizes these statistics for the LDA models. This shows that LDA has reasonable accuracy when applied to the selected subchronic attributes, but applying it to the structural alert attributes has accuracy just barely better than random.

		Predicted Class		
		Negative	Positive	Total
Actual Class	Negative	53	8	61
	Positive	32	81	113
	Total	85	89	174

Table 7: Cross-Validated Results of LDA Combined

Class	Test Agents With No Subchronic Present	Test Agents With With Subchronic Present	Overall
Negative Predictivity	57%	67%	62%
Positive Predictivity	70%	95%	91%
Specificity	84%	89%	87%
Sensitivity	30%	82%	72%
Overall Accuracy	58%	84%	77%

Table 8: Accuracy of LDA Models

### 3.2 Predictive Value of Bayesian Classifiers

Table 9 combines the results of cross-validating the two Bayesian classifiers to measure their predictive capabilities in general. Table 10 summarizes a number of statistics for these models. The fact that positive predictivity is high for both of these networks means that, if a network predicts that a test agent is carcinogenic, it is likely that the prediction is correct. This is equivalent to saying that the likelihood of a false positive prediction is quite low. Sensitivity is also quite high for test agents which exhibited at least one of the selected subchronic attributes, so for that group, the likelihood of a false negative is relatively low. However, sensitivity is lower for test agents which exhibited none of the selected subchronic attributes, so the likelihood of a false negative for that group of test agents is much higher.

		Predicted Class		
		Negative	Positive	Total
Actual Class	Negative	60	4	64
	Positive	21	106	127
	Total	81	110	191

Table 9: Cross-Validated Results of Bayesian Classifier Combined

## 4 Discussion

While the cross-validation predicted accuracies for these models is competitive with other automated predictors in this field and with human experts, the off-training set accuracy

Class	Test Agents With No Subchronic Present	Test Agents With With Subchronic Present	Overall
Negative Predictivity	64%	82%	74%
Positive Predictivity	100%	96%	96%
Specificity	100%	90%	94%
Sensitivity	77%	92%	83%
Overall Accuracy	72%	92%	87%

Table 10: Accuracy of Bayesian Classifiers

does not appear to be as high as predicted. We predicted the 30 chemicals in the PTE-2 competition sponsored by NIEHS. PTE-2 consisted of 30 chemicals which were in the process of being evaluated by NTP. The goal of the experiment was to measure prediction algorithms by having them make predictions before the result was known. We used our models to make predictions for all 30 of the PTE-3 chemicals. After the first 9 were classified by NTP, our accuracy were just over 50% (we had 5 out of the 9 correct). It is possible that this is a statistical anomaly and that our overall performance will be better than the performance on those nine chemicals, but these results gave us reason to re-consider our approach.

All of machine learning depends on the training set being rich enough to represent all of the population being modeled. Since we only have 226 chemicals in the training set, it is doubtful that they fully represent all chemicals. The question is whether they represent the set of chemicals that NTP is going to test for carcinogenicity. It is our opinion that they do not represent a rich enough training set for this problem. We believe that the reason our off-training set results do not match our predicted accuracy is that our training set does not fully represent the population. Since the number of chemicals studied by NTP is limited, there is no way to add more chemicals to the data set.

Having realized the weakness in the data set, the problem becomes how to proceed in trying to make accurate predictions. First, the linear discriminant analysis requires normality assumptions which are unlikely to be valid for this application, so we would like to learn (not necessarily naive) Bayesian classifiers directly from the data. Second, a number of the sub-chronic attributes are extremely rarely positive (positive in only 1 or 2 chemicals). It is our opinion that the linear discriminant analysis and other learning techniques are seizing on these attributes because of their high predictivity. Essentially, the learning is generalizing one or two instances across the entire population, which is unlikely to be valid. Therefore,

in our most recent research we discard all attributes which are positive in fewer than three chemicals.

Finally, the end-point we are modeling (positive or negative carcinogenicity) is really a surrogate for many end-points for many different types of carcinogenicity. We have in effect been asking the learning algorithm to model many different pathways to many different types of cancer in many different organs. In hindsight, the odds of a learning algorithm being able to accomplish our goal with only 226 examples are very small. However, NTP maintains information about exactly what cancers were associated with each chemical. Therefore, we plan to experiment with changing the end-point to predict a specific type of cancer at a specific organ. We believe that this may lead to models which give some insight into biological pathways and may allow our small training set to result in reasonably accurate models.

## Acknowledgements

Special thanks is owed to Dr. Ann Richard and Mr. Phillip Boone of U.S. EPA for gathering the physical chemical data used in the training and test sets. This work was funded in part by the Laboratory for Environmental Carcinogenesis and Mutagenesis of the National Institute of Environmental Health Sciences, NIH.

## References

- [1] Ashby, J., Tennant, R.W., Zeiger, E. and Stasiewicz, S. Classification according to chemical structure, mutagenicity to salmonella and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. *Mutation Research* 223:73-101 (1989).
- [2] Bakale G and McCreary RD. Prospective  $k_e$  screening of potential carcinogens being tested in rodent bioassay by the U. S. National Toxicology Program. *Mutagenesis* 7(2): 91-94 (1992).
- [3] Dougherty James, Kohavi Ron, and Sahami Mehran. Supervised and Unsupervised Discretization of Continuous Features. *Machine Learning* 12:194-202 (1995).
- [4] Fayyad Usama M, Irani Keki B. Multi-interval discretization of continuous-valued attributes for classification learning. In: *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, August 1995, Morgan Kaufmann:1022-1027 (1995).

- [5] Huff James, McConnell EE, and Moore JA. The National Toxicology Program Toxicology Data Evaluation Techniques and Long-Term Carcinogenesis Studies. Safety Evaluation of Drugs and Chemicals: 441-447 (1985).
- [6] Jensen FV, Olesen KG, Andersen SK. An Algebra of Bayesian Belief Universes for Knowledge-Bases Systems. Networks 20:637-659 (1990).
- [7] Johnson Richard A, Wichern Dean W. Applied Multivariate Statistical Analysis. Englewood Cliffs, New Jersey: Prentice Hall, 1992.
- [8] Lauritzen SL, and Spiegelhalter DJ, Local computations with probabilities on graphical structures and their application to expert systems (with discussion) Journal of the Royal Statistical Society B 50:157-224 (1988).
- [9] Pearl J. Probabilistic Reasoning in Intelligent Systems. San Francisco, CA, :Morgan Kaufmann Publishers, Inc., 1988.
- [10] Peot Mark A and Shachter Ross D. Fusion and propagation with multiple observations in belief networks. Artificial Intelligence 46:299-318 (1991).